mathematics

MDPI

Article

Intelligent Visual Detection of Surrounding Personnel for Underground Load–Haul–Dump Vehicles in Complex Mining Environments

Pingan Peng [1,2], Kaixuan Cheng [1,*], Chaowei Zhang [1], Xuhe Li [1], Haoyue Zhang [1] and Shuangwei Gong [1]

[1] School of Resources and Safety Engineering, Central South University, Changsha 410083, China; ping_an@csu.edu.cn (P.P.); 235511057@csu.edu.cn (C.Z.); 245512095@csu.edu.cn (X.L.); 245512088@csu.edu.cn (H.Z.); 245512114@csu.edu.cn (S.G.)
[2] State Key Laboratory for Fine Exploration and Intelligent Development of Coal Resources, Xuzhou 221116, China
* Correspondence: 245511157@csu.edu.cn

Abstract

In underground mining environments, collisions between Load–Haul–Dump (LHD) machines and personnel pose serious safety risks. To address challenges such as uneven illumination, dust interference, occlusion, and limited computational resources, this paper proposes FP-DETR, an improved RT-DETR framework for underground personnel detection. The model integrates a CSP-FFCM module for efficient spatial–frequency feature extraction, a polarity-aware linear attention-based POTE module for enhanced global feature interaction, and a Matching-Aware Loss to improve confidence–quality alignment and reduce false alarms. Experiments on a self-constructed dataset show that FP-DETR achieves 88.5% mAP@0.5 and 58.6% mAP@0.5:0.95, improving RT-DETR-r18 by 1.5% and 3.5%, respectively, while reducing parameters from 20.0 M to 15.8 M and GFLOPs from 57.3 to 51.2. Field tests show a reduction in frame-level FPR from 4.1% to 3.4%. However, validation is limited to no-person scenarios, and full operational evaluation is required for comprehensive assessment.

Keywords: underground personnel detection; RT-DETR; Fourier Convolution; polarity-aware attention; intelligent mining

MSC: 68T75; 68T07

1. Introduction

Underground mining is one of the most hazardous industrial activities, where harsh environmental conditions like limited visibility and dark tunnels significantly increase the risk of accidents. In underground metal mines, Load–Haul–Dump (LHD) machines play a critical role in transporting ore and materials, and their operational efficiency and level of intelligence directly impact the mine's overall productivity. However, LHDs often operate in narrow, dark tunnels with blind spots, making them highly susceptible to collisions with nearby personnel, which can lead to equipment damage and severe casualties [1]. Therefore, exploring how to use intelligent technology to reduce these accidents by accurately and reliably detecting personnel around LHDs in real time has become a pressing challenge to enhance the intelligence of mine safety management [2].

As shown in Figure 1, the underground environment is distinct from surface conditions, presenting several unique challenges to safe operations. These include prolonged

Academic Editor: Oleg Sergiyenko

Received: 7 May 2026
Revised: 2 June 2026
Accepted: 4 June 2026
Published: 5 June 2026

low or uneven lighting, high concentrations of airborne dust, high humidity, and confined, winding tunnel spaces. Moreover, underground production involves the frequent movement of large machinery and personnel, creating complex scenarios where equipment and human activity areas overlap. These combined factors of harsh environments and complex operational modes significantly increase the overall risk of underground mining.

(**a**) (**b**)

Figure 1. The underground production environment. (**a**) Uneven illumination; (**b**) Narrow tunnels.

For example, low light and dust severely reduce visibility, interfering with a person's visual perception and operational behavior. This significantly degrades situational awareness for both workers and operators and for equipment operators to fully perceive nearby individuals or obstacles. As shown in Figure 2, LHDs are large vehicles with significant blind spots during operation. The constricted nature of subterranean tunnels, compounded by the frequent convergence of heavy machinery and personnel in shared spaces, substantially diminishes safety margins and elevates the likelihood of collision incidents.

Figure 2. Load–Haul–Dump vehicles.

Traditional safety management measures and human-based monitoring are often inadequate for handling these complex, dynamic underground environments and high-intensity operational demands. While various personnel detection technologies have been applied in mining safety, existing solutions still face numerous challenges and limitations [3,4]. Furthermore, current research often focuses on improving detection accuracy while overlooking the critical need for a low false alarm rate. Frequent false alarms not only disrupt normal production and reduce efficiency but can also lead operators to distrust the system, causing them to ignore genuine warning signals [5].

To address these challenges and advance underground personnel detection technology, this paper proposes a new detection method, FP-DETR, based on RT-DETR. Our method aims to reduce the false alarm rate, improve detection accuracy, and decrease model parameters. The proposed method integrates Fast Fourier Convolution (FFC) and Polarity-aware Transformer Encoder. First, FFC is introduced into the backbone network to fuse multi-scale spatial features through spatial–frequency domain convolution. This allows for more efficient extraction of global information, effectively tackling issues such as occluded personnel, uneven lighting, and dust-related noise in complex underground environments, thereby improving the detection of personnel around LHDs. Second, a Polarity-aware Transformer Encoder mechanism is incorporated into the neck network. This mechanism is designed to address the high computational complexity of Transformer self-attention and the limited representational power of traditional linear attention in underground settings. By explicitly modeling the interactions of same-signed and opposite-signed query–key pairs, it enhances the model's discriminative power and efficiency.

Furthermore, to address the data imbalance between personnel and background, which often leads to false positives, we introduce the Matching-Aware Loss (MAL). A high-quality field dataset was collected and annotated specifically for the underground environment. This dataset was further augmented with pedestrian pairs from the RTTS dataset, and underground background images were included as negative samples during training. This approach is designed to reduce false alarms and enhance the model's robustness in real-world scenarios. The FP-DETR algorithm for detecting personnel around underground LHDs demonstrates potential applicability importance for ensuring the safety of miners and enhancing overall mine safety production levels.

2. Related Work

According to their detection principles and application scenarios, common traditional personnel detection and positioning systems include technologies such as GPS, UWB, RFID, and LIDAR. However, the unique and challenging environmental conditions found in underground mines, including dust, poor visibility, high humidity, and elevated temperatures, severely hinder the use of these sensors. While the Global Positioning System (GPS) offers high precision for surface operations, its dependence on satellite signals makes it unusable in subterranean environments [6]. Ultra-wideband (UWB) radar, which detects obstacles by transmitting electromagnetic waves, is often prone to false positives and struggles to differentiate between people and objects [7]. RFID (Radio-Frequency Identification) systems use tags worn by workers, which are detected by readers in hazardous zones to trigger visual and auditory alerts for vehicle operators. Yet, these systems cannot provide a precise location for employees, require every person and vehicle to be tagged, and suffer from signal attenuation caused by rough walls [8]. Lastly, LIDAR-based positioning systems can quickly detect the distance between all objects within a scanning range of up to 50 m in enclosed underground spaces; however, they lack the ability to accurately distinguish between people and objects, resulting in significant positioning errors [9].

Over the last few years, artificial intelligence, particularly computer vision, has rapidly and profoundly empowered various industries. Computer vision-based anti-collision systems have already been successfully applied in open-pit mines, quarries, and similar environments [10]. Jing Zhang proposed a crowd-sensing technology based on the intelligent industrial Internet of Things for safety monitoring in coal mines, which has achieved good results [11]. MSC-DETR is a lightweight real-time foreign object detection model based on RT-DETR, specifically designed for underground coal mine belt conveyors under challenging conditions such as uneven illumination and heavy background interference. It achieves strong performance by introducing MAFEA, SEFFN-AIFI, and CEAFPN modules,

reducing parameters and GFLOPs by 55.2% and 70.2% respectively while attaining 93.3% mAP50 at 124.6 FPS on the CUMT-BelT dataset. However, its generalization capability across different mine environments and robustness under severe dust or extreme occlusion remain to be further validated [12]. Gang Wang optimized the YOLOv8 model for small targets on drones and proposed a target detection model based on drone aerial photography scenarios, called UAV-YOLOv8, which achieves a good balance between detection performance and resource consumption [13]. Due to its high efficiency and adaptability, computer vision has become a hot topic in underground personnel detection research, offering an effective way to reduce accidents and fatalities in subterranean mining operations [14–16]. Several studies have contributed to this field: Wei proposed the PftNet detection network, a novel detector that outperforms single-stage methods while maintaining the accuracy of two-stage approaches, though its identification module involves a cumbersome multi-step process [17]. Wang et al. introduced a real-time underground obstacle detection algorithm based on DeblurGANv2 and an improved YOLOv4, but its detection accuracy in complex conditions needs improvement, and the network structure has room for optimization [18]. Shao's new Rep-YOLO (re-parameterized YOLO) detection algorithm improved the model's recall, AP50, and FPS [19]. Wang et al. proposed a lightweight network based on an improved YOLOv5 for multi-object detection in underground rail locomotive driving areas; their results showed a 12.3% reduction in parameters and a 1.7% increase in mAP, making it more suitable for multi-target detection in these specific environments [20]. Tian Z. showed that data augmentation on low-light underground images before using YOLOv8 improved detection performance [21]. Feng Tian proposed an improved Retinex algorithm for low-light mine image enhancement. Experimental data show that the algorithm improves image brightness, prevents halo artifacts while preserving edge details, and reduces the impact of noise. However, image enhancement requires additional computing power and is not suitable for the real-time requirements of downhole detection [22]. Ni et al. used an improved YOLOv8 algorithm combined with the ray method to determine if underground personnel had entered a dangerous zone, demonstrating higher accuracy than the standard YOLOv8 model in handling underground environmental variations [23]. Jin et al. presented an improved YOLOv8-based method to address issues of high computational consumption and slow detection speed, although their research was not sufficiently thorough regarding more complex low-light conditions and false positive rates (FPR) [24]. Lastly, Yadong Li integrated an efficient channel attention (ECA) mechanism into the YOLOv11 model, enhancing its ability to focus on key features and significantly improving detection accuracy [25].

Significant progress has been made in computer vision-based underground personnel detection systems. However, the inherent limitations of the receptive field in convolution operations and the inductive bias of convolutional structures make it challenging to establish long-range dependencies and global context. Moreover, algorithms like the YOLO series often require non-maximum suppression (NMS) as a post-processing step to remove duplicate bounding boxes, which increases computational complexity and hinders true end-to-end detection performance.

The underground environment is complex with poor lighting conditions, and underground personnel targets are often small and have indistinct features. The local modeling of convolutional neural networks (CNNs) is more susceptible to factors such as illumination, dust interference, and occlusion. Furthermore, because underground personnel are often occluded, small, and blurry, CNNs lack global modeling capabilities, suggesting that CNN-based models may not be well-suited for personnel detection in underground environments.

Existing research primarily focuses on improving detection accuracy in known scenes, neglecting issues faced in practical applications in complex underground scenarios, such as large model parameters, limited generalization ability, and the generation of numerous false positives that interfere with normal production operations. Therefore, there is a need to develop personnel detection methods specifically suited for the unique underground environment.

Unlike CNNs, Transformers possess the ability to capture complex spatial variations and establish global relationships between image elements through their self-attention mechanism [26]. The exceptional generalization and robustness of Transformer-based models have been demonstrated in previous studies. Carion et al. proposed the Detection Transformer (DETR), which integrates Transformers into object detection by using a CNN as a backbone and combining it with a Transformer encoder–decoder network [27]. Compared to single-stage detectors (e.g., YOLO series) and two-stage detectors (e.g., Faster R-CNN) that are based on CNNs, DETR offers a fully end-to-end trainable architecture. This eliminates the need for post-processing steps like Non-Maximum Suppression (NMS), as well as the need for prior knowledge and constraints, which significantly simplifies the object detection pipeline. Furthermore, DETR has shown superior performance in the field of object detection when compared to its CNN-based counterparts. RT-DETR, proposed by Baidu, is an end-to-end object detection framework designed to achieve faster inference speeds while maintaining high detection accuracy [28]. RT-DETR employs an efficient hybrid encoder that integrates multi-scale feature interactions, allowing for improved detection speed while preserving high accuracy. Key improvements of RT-DETR include the following: an efficient hybrid encoder is used to reduce computational redundancy; a method called uncertainty-minimal query selection optimizes target queries to enhance detection accuracy; and the model supports flexible speed adjustments to adapt to various real-time detection requirements.

To our knowledge, there is limited research on using end-to-end DETR models for underground personnel detection and recognition compared to the YOLO series. An enhanced version of RT-DETR, named RT-DETR-MS, was proposed by Song et al. to improve personnel detection in dense scenes, and the results showed an improved ability to detect small objects, but it was not applicable to underground environments [29]. Tian's improved PDP-RTDETR model achieved a mean average precision (mAP) of 56.6% for small object detection on a self-built dataset, successfully improving small object detection accuracy in mine scenarios [30]. RT-DETR surpasses the YOLO series of comparable sizes in both speed and accuracy on the COCO dataset. Therefore, it is feasible to use RT-DETR as a foundational model to explore personnel detection around LHD in complex environments.

3. Datasets

Constructing and pre-processing high-quality datasets are crucial steps for training effective object detection models. While mainstream pedestrian detection datasets such as CityPersons [31], Caltech Pedestrian [32], and COCO [33] possess a certain level of diversity and complexity, they primarily focus on urban ground-level scenes. These datasets fail to encompass the unique visual interferences characteristic of underground environments and using them directly for training leads to a lack of generalization ability when the model is deployed in real-world scenarios. For the specific task of personnel detection around underground LHD, a dataset must genuinely reflect the complexity of the working environment. Therefore, building a representative and task-specific dataset is essential for improving detection accuracy. To align with the actual detection requirements of the underground environment, the dataset used in this paper is composed of self-collected

images of underground miners and images with "Person" labels extracted from RTTS data [34].

3.1. Dataset Collection

The proprietary underground personnel dataset used in this study was collected from two underground mining sites in China: The Luohe Iron Mine in Anhui Province and the Panlong Lead-Zinc Mine in Guangxi Province. The dataset contains diverse underground scenes acquired under real operating conditions, including low illumination, dust interference, occlusion, uneven exposure, and complex equipment backgrounds. As illustrated in Figure 3, the collected images cover multiple working scenarios involving underground personnel near LHDs, mining carts, tunnel intersections, and curved roadways, thereby ensuring strong representativeness for underground personnel detection tasks.

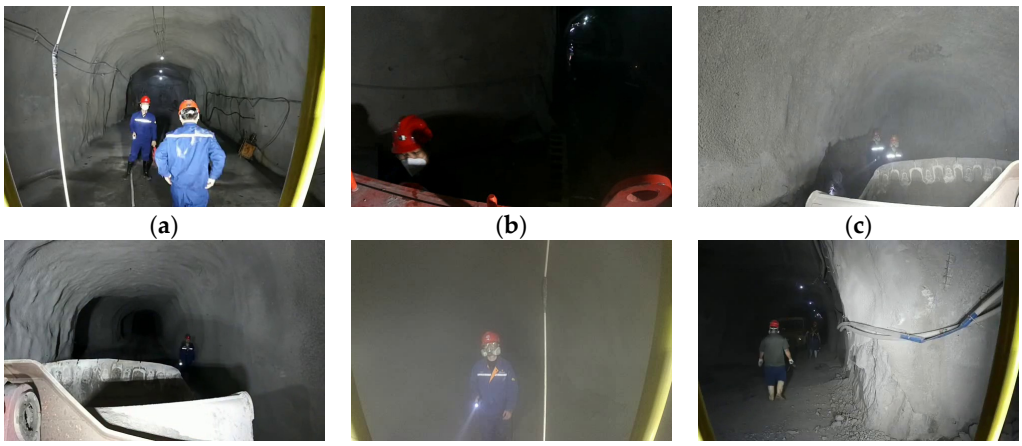


Figure 3. Data collected from different underground scenes. (**a**) Bright lighting conditions; (**b**) Low-light conditions; (**c**) Highly occluded scene; (**d**) Long-distance scene; (**e**) High-dust environment; (**f**) Similar background scene.

The first data source was the Luohe Iron Mine located in Hefei, Anhui, China. Data collection was conducted from 1 October to 20 October 2024. During this period, five OAK-D-Pro-W-PoE cameras (Luxonis, Broomfield, CO, USA) were mounted on a Sandvik LH514E load-haul-dump (LHD) vehicle (Sandvik Mining and Rock Solutions, Stockholm, Sweden) to continuously capture real-time video streams during underground operations, enabling comprehensive coverage of the surrounding working area. The second data source was the Panlong Lead-Zinc Mine in Guilin, Guangxi, China, where data were collected from 10 March to 12 March 2025. In this mine, underground personnel videos were manually captured under different lighting conditions, shooting angles, and occlusion scenarios.

The collected videos were first processed by frame extraction using the OpenCV (version 4.12.0, Open Source Computer Vision Library) for video frame extraction. To avoid excessive redundancy caused by adjacent video frames, one frame was sampled every 30–60 frames. Subsequently, blurry frames, ghosting frames, and low-quality samples were manually removed. Frames containing underground personnel were retained as positive samples, while tunnel scenes without personnel were retained as negative samples.

To address the concern of potential data leakage caused by frame-level random splitting, a strict video-level data partition strategy was adopted. Specifically, all frames extracted from the same video sequence were assigned exclusively to either the training set or the validation set. Therefore, highly similar frames from the same video could not simultaneously appear in both subsets. This strategy effectively prevents overestimation of model performance and improves the reliability of the evaluation results.

Due to the harsh underground environment, including dust interference, low illumination, and weak texture features of miners' clothing, personnel appear relatively sparsely in practical underground scenarios. To better reflect the real distribution characteristics of underground environments and reduce false positive detections, 2000 negative images without personnel were additionally collected from tunnel scenes. These negative samples were also included in both the training and validation sets following the same 8:2 split ratio.

In addition to the self-collected underground dataset, the Real-world Task-driven Testing Set (RTTS) was introduced to further improve dataset diversity and model robustness. RTTS contains 4322 real-world hazy images with annotations for five object categories, including person, bicycle, bus, car, and motorbike. As shown in Figure 4, the RTTS dataset exhibits visual characteristics like underground environments, such as low contrast, blur, haze-like interference, and reduced visibility. Therefore, 2734 images annotated with the "person" category were extracted and incorporated into the dataset to improve the detector's adaptability to low-visibility and degraded underground conditions.

Figure 4. Pedestrians in different scenes within the RTTS dataset. (**a**) Foggy scenes; (**b**) Bright scenes; (**c**) Low-contrast scenes; (**d**) Complex lighting conditions.

3.2. Data Augmentation

Due to the difficulty and high cost of collecting underground personnel images, the amount of available training data was relatively limited. To improve dataset diversity and enhance the generalization capability of the detector, the self-collected underground dataset was first divided into training and validation subsets using the video-level splitting strategy described in Section 3.1 with an 8:2 ratio. Data augmentation was subsequently applied exclusively to the training subset of the self-collected underground dataset using the Albumentations library (version 2.0.5), while the validation subset retained only original non-augmented images. The RTTS dataset was not augmented in this study. Therefore, augmented versions derived from the same original frame could not simultaneously appear in both the training and validation subsets, thereby ensuring a fair and reliable model evaluation process.

As illustrated in Figure 5, both geometric and photometric augmentation strategies were adopted to simulate the complex visual characteristics of underground environ-

ments, including viewpoint variation, illumination fluctuation, blur, dust interference, and low visibility.

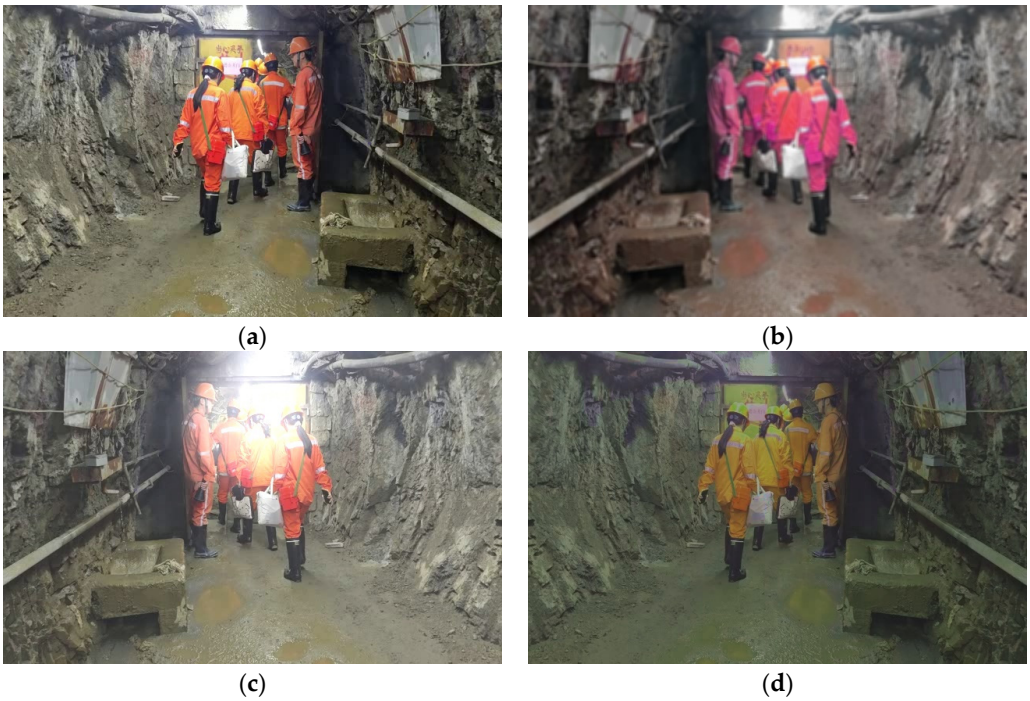


Figure 5. Original image and images after various data augmentation techniques using the Albumentations library (version 2.0.5). (**a**) Original image; (**b**) Horizontal flipping and color variation; (**c**) Horizontal flipping and local exposure; (**d**) Color variation.

For geometric augmentation, affine transformation was applied to simulate variations in camera viewpoints and personnel postures during underground operations. The affine transformation included random scaling, translation, rotation, and shear operations, where the scaling factor ranged from 0.5 to 1.5 and the shear angle ranged from $-45°$ to $45°$. In addition, perspective transformation, elastic deformation, grid distortion, and random cropping were introduced with relatively low probabilities to enhance the robustness of the model to geometric deformation and partial occlusion.

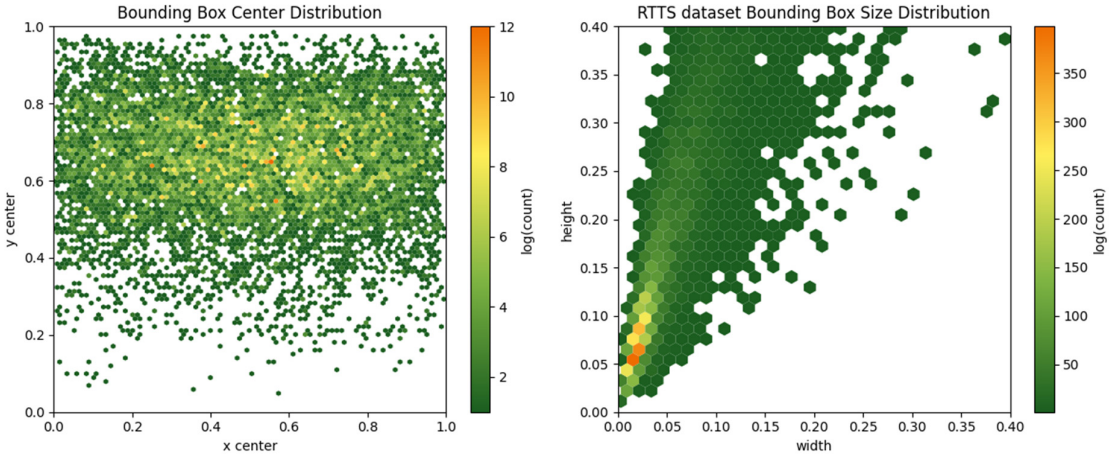
To simulate image degradation commonly observed in underground environments, several photometric augmentations were further applied, including Gaussian noise, image compression, random brightness and contrast adjustment, fog simulation, shadow simulation, and grayscale conversion. Among them, JPEG compression quality was randomly adjusted within the range of 50–100 to imitate image quality degradation caused by video transmission and acquisition devices. Random brightness and contrast transformations were employed to simulate uneven underground illumination conditions, while fog and shadow augmentations were used to reproduce dust interference and local occlusion effects frequently encountered in underground scenes.

To ensure annotation consistency during augmentation, all bounding boxes were synchronously transformed in YOLO format. Bounding boxes with visibility lower than 10% after augmentation were automatically discarded to avoid generating unreliable annotations. Through the combined application of these augmentation methods, the number of underground personnel training samples was substantially increased after augmentation, thereby improving dataset diversity.

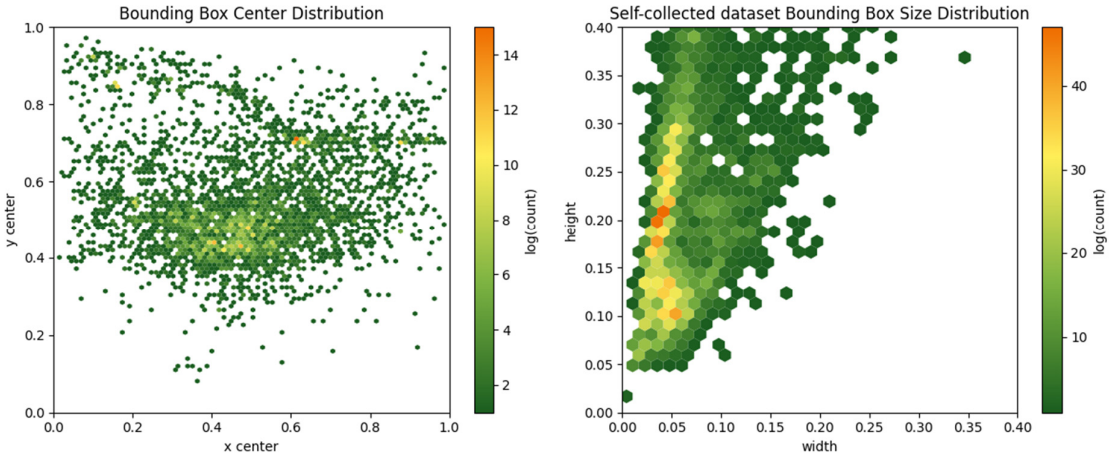
3.3. Dataset Merging and Analysis

To improve the robustness and generalization ability of the detector, the augmented underground dataset was merged with the filtered RTTS personnel dataset for hybrid training. Before merging, a quantitative analysis was conducted to evaluate the consistency between the two datasets and reduce the risk of domain shift.

First, the spatial distribution and scale characteristics of human targets were analyzed. As shown in Figure 6, the bounding boxes in both datasets exhibit similar spatial clustering patterns in the normalized image coordinate system. Most human targets are concentrated in the middle and lower regions of the images, which is consistent with the camera installation perspective on underground LHD equipment and the typical activity regions of miners. In terms of target scale distribution, both datasets show consistency in normalized bounding box width and height. Most targets are concentrated within the normalized size range of 0.05–0.20, indicating that both datasets predominantly contain small-scale pedestrian targets. This consistency suggests that the RTTS dataset shares similar geometric characteristics with underground personnel images, thereby supporting the feasibility of dataset fusion.



(**a**)



(**b**)

Figure 6. Target size and position of people in RTTS dataset and self-collected dataset. (**a**) RTTS dataset bounding box position and size; (**b**) Self-collected dataset bounding box position and size.

To further evaluate the compatibility between the two datasets, pixel-level distribution consistency analysis was performed using Jensen–Shannon (JS) divergence, Kullback–

Leibler (KL) divergence, and Wasserstein distance across RGB channels. The statistical results are presented in Table 1.

Table 1. Pixel distribution difference metric.

Metric	R Channel	G Channel	B Channel
JS Divergence	0.0161	0.0153	0.0141
KL Divergence	0.0610	0.0578	0.0539
Wasserstein Distance	20.54	19.93	20.03

As shown in Table 1, the JS divergence values of all RGB channels are approximately 0.015, indicating a certain degree of similar distribution shapes between the two datasets. Meanwhile, the KL divergence values remain below 0.07, suggesting only minor information discrepancy. Although the Wasserstein distance values are approximately 20, reflecting slight global illumination differences, the overall deviation remains relatively small compared with the full pixel intensity range of 0–255.

Furthermore, as illustrated in Figure 7, the RGB pixel distributions of the underground dataset and the RTTS dataset exhibit highly consistent overall trends across all three channels. The distribution curves almost overlap within the medium-intensity regions where most pixels are concentrated, indicating strong consistency in low-level visual characteristics such as illumination patterns, color composition, and background appearance. Slight deviations appear mainly in high-intensity regions, where the underground dataset exhibits slightly brighter pixel values due to differences in underground lighting equipment and imaging conditions. Nevertheless, the overall distribution similarity remains high, demonstrating that the RTTS dataset can be effectively integrated without introducing severe domain bias.

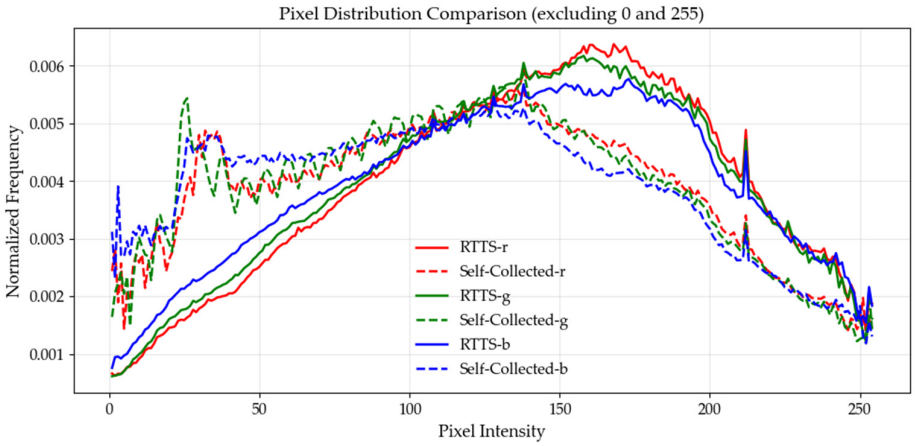


Figure 7. Comparison of RGB Pixel Distributions between the Self-Collected Dataset and RTTS dataset.

Figure 8 shows the location and size of the target bounding boxes in the merged dataset. Since the RTTS dataset does not provide annotations in YOLO format, all annotations were first converted into normalized YOLO-format labels. Subsequently, only the "person" category was retained, while all other object categories were removed. The self-collected underground dataset was annotated using LabelImg (version 1.8.6), and all annotations were manually verified to ensure labeling accuracy and consistency. To improve annotation reliability and consistency, all dataset annotations were independently reviewed and cross-validated by two annotators. In cases where annotation discrepancies occurred, the corresponding samples were jointly reviewed and discussed by the

two annotators until a consensus was reached before final dataset construction. After dataset merging, the final dataset contained 5930 positive samples with personnel annotations, including 3196 augmented underground positive samples and 2734 RTTS personnel samples. In addition, 2000 underground negative samples without personnel were retained. The ratio between positive and negative samples was approximately 2.96:1, which is consistent with the sparse personnel distribution characteristics commonly observed in underground mining environments.

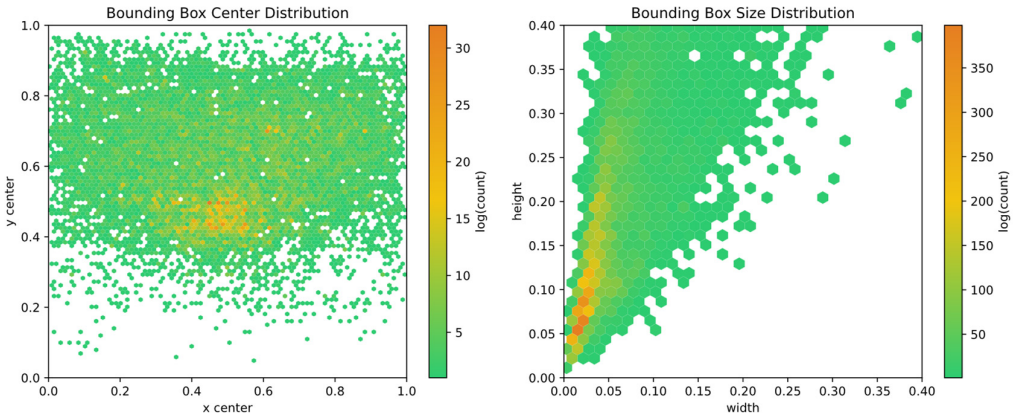


Figure 8. Visualization and distribution of object bounding boxes in the merged dataset.

The detailed statistical information of the final merged dataset is summarized in Table 2. Following the video-level splitting strategy described in Section 3.1, the self-collected underground dataset was first divided into training and validation subsets using an 8:2 ratio. Data augmentation was applied exclusively to the underground training subset, while the validation subset retained only original non-augmented images. Subsequently, the underground training subset and the corresponding RTTS training subset were merged to construct the final training dataset. The final validation dataset consisted solely of original non-augmented images from both datasets. Both positive and negative samples were included throughout the dataset construction and partitioning process.

Table 2. Merged Dataset Details.

Dataset Source	Original Positive Samples	Final Training Samples	Validation Samples	Negative Samples
Luohe Iron Mine	812	1449	363	1000
Panlong Lead-Zinc Mine	1188	1747	441	1000
RTTS	2734	2189	545	0
Total	4734	5385	1349	2000

4. Methodology

The performance of RT-DETR in underground personnel detection is largely determined by the representation capability of its backbone network and the effectiveness of its feature interaction mechanism. However, the original RT-DETR framework faces three major limitations in underground mining environments. First, the ResNet-18 backbone mainly relies on local convolutional operations, which limits its ability to capture discriminative features under conditions involving uneven illumination, dust interference, low contrast, and target occlusion. Second, the standard self-attention mechanism adopted in the Attention-based Intra-scale Feature Interaction (AIFI) module introduces quadratic computational complexity, making it difficult to satisfy the real-time requirements of personnel detection around Load–Haul–Dump (LHD) vehicles. Moreover, conventional self-attention insufficiently models directional and local spatial dependencies that are

critical for identifying partially occluded or low-visibility personnel. Third, the Varifocal Loss (VFL) used in RT-DETR is sensitive to severe foreground–background imbalance, which frequently results in unstable confidence estimation and excessive false positives in underground scenes.

To address these challenges, an improved RT-DETR framework termed FP-DETR is proposed, as illustrated in Figure 9. The proposed framework introduces three key modifications. First, a Cross-Stage Partial Fused Fourier-Conv Mixer (CSP-FFCM) module is designed to enhance spatial–frequency feature representation and improve robustness against illumination degradation and environmental interference. Second, a POTE is introduced to replace the original AIFI module, enabling efficient global dependency modeling with linear computational complexity while preserving directional feature sensitivity. Finally, a Matching-Aware Loss (MAL) is adopted to alleviate the mismatch between positive and negative samples and improve the reliability of confidence estimation in challenging underground environments.

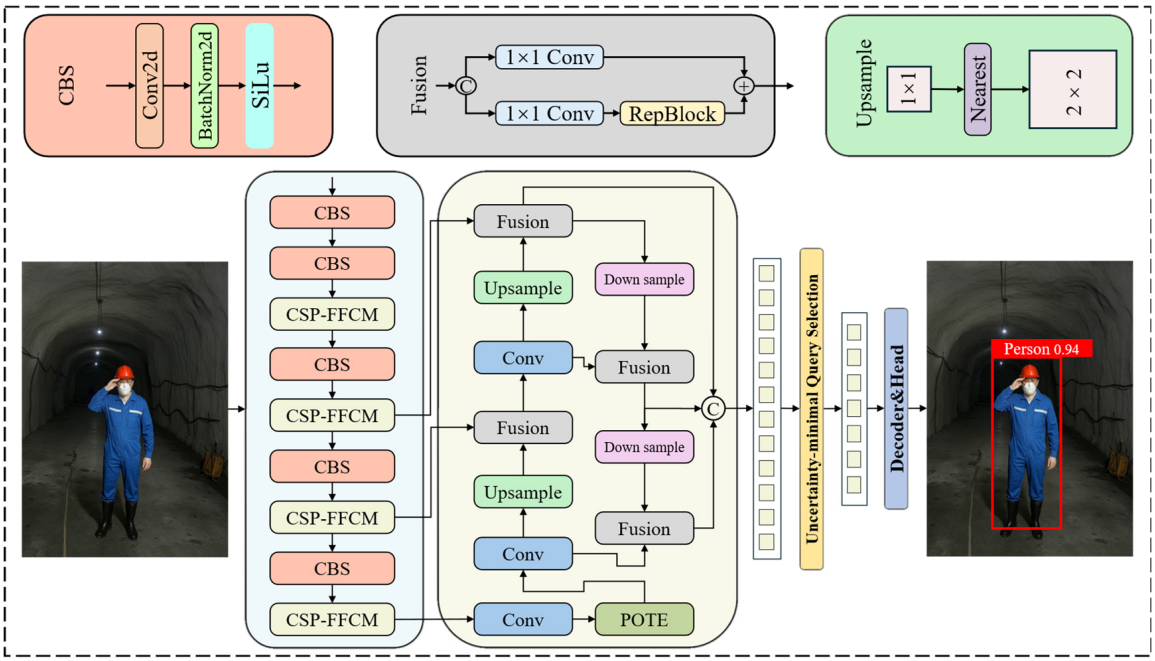


Figure 9. Overall architecture of FP-DETR. "©" denotes the concatenation (Concat) operation in the network architecture, "⊕" denotes element-wise addition (Add), and red bounding boxes indicate the detected objects.

4.1. CSP-FFCM Module

Conventional convolutional backbones mainly extract features through local receptive fields, which limits their ability to model long-range dependencies and global structural information. In underground mining environments, severe illumination fluctuations, dust interference, and low-contrast targets often weaken local texture cues, resulting in degraded feature discrimination. To improve feature representation under these conditions, a Cross-Stage Partial Fused Fourier-Conv Mixer (CSP-FFCM) module is introduced to jointly exploit spatial-domain and frequency-domain information.

As shown in Figure 10, the CSP-FFCM module's architecture draws inspiration from the Cross Stage Partial (CSP) concept [35], leveraging the cross-fusion of shallow and deep features to enhance the diversity of feature representation. Unlike the traditional CSP module, we replaced the original Bottleneck structure with a novel Fused Fourier-Conv Mixer (FFCM) operation. This module combines the global modeling capabilities of the Fourier transform with the local perception characteristics of a convolutional neural

network, enabling complementary modeling in both the frequency and spatial domains [36]. Previous research by Gao et al. has demonstrated the effectiveness of the FFCM in image denoising, specifically for removing rain [37], which further supports its ability to extract enhanced features through spatial–frequency domain convolution.

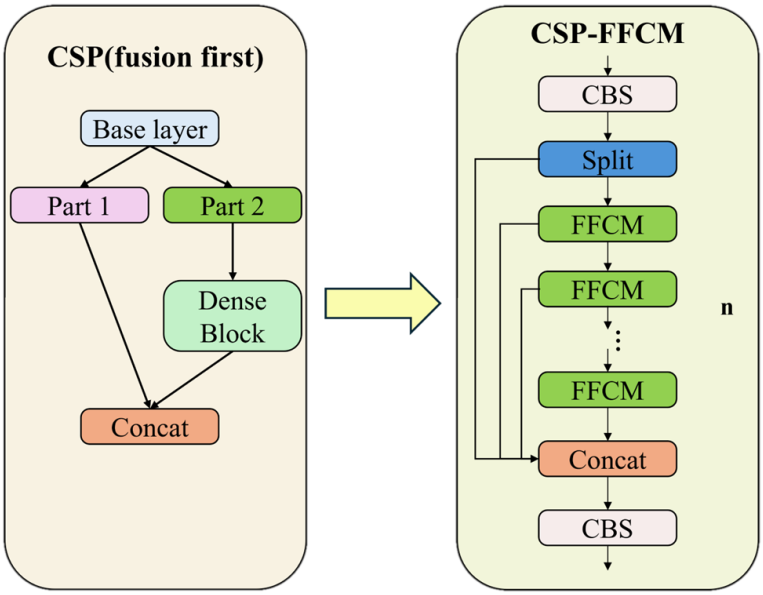


Figure 10. A schematic diagram of the traditional CSP structure and the proposed CSP-FFCM structure with the introduction of the FFCM module. The arrows indicate the integration of the CSP design philosophy with the FFCM module to construct the proposed CSP-FFCM structure.

As shown in Figure 11, FFCM is composed of several subcomponents. First, the input features are expanded from dim to 2 × dim dimensions using a 1 × 1 convolution for inter-channel information interaction. The features are then split along the channel dimension and processed by a 3 × 3 depthwise convolution and a 5 × 5 depthwise convolution, respectively. This process captures fundamental features like edges and textures, and the multi-scale feature extraction helps maintain stable detection performance under non-uniform lighting conditions. Following this, the module enters the Fourier Unit frequency-domain processing branch, which transforms the feature map into the frequency domain. Through specific frequency transformations and filtering, this branch highlights crucial frequency components, such as those corresponding to the contours of personnel, thereby enriching the feature hierarchy and semantic information. Finally, the features from the convolutional branch and the frequency-domain branch are fused using a weighted summation strategy, where the weight parameters are learned during training. This approach dynamically balances the contribution of both feature types to generate a more representative output feature map, providing a higher-quality foundation for subsequent detection tasks.

The basic computational process of the FFCM can be described by the following formula: First, the input features undergo channel expansion and grouped convolution to extract local information.

Given an input feature tensor $X \in \mathbb{R}^{B \times C \times H \times W}$, the proposed Fused Fourier Convolution Mixer (FFCM) aims to jointly capture local spatial structures and global spectral dependencies through a coupled spectral-spatial representation framework.

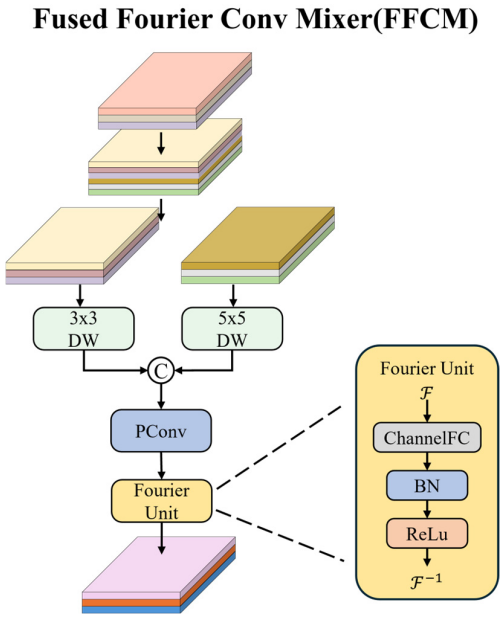


Figure 11. Detailed Architecture Diagram of Fused Fourier Conv Mixer. The dashed line indicates an enlarged detailed view of the corresponding module, and © denotes the concatenation (Concat) operation.

First, the input feature is projected and decomposed into two complementary subspaces for multi-scale local modeling:

$$\mathrm{X}_l = [\mathcal{D}_3(\psi_1(\mathrm{X})), \mathcal{D}_5(\psi_2(\mathrm{X}))] \tag{1}$$

where $\psi_i(\cdot)$ denotes channel projection and splitting operations, while $\mathcal{D}_k(\cdot)$ represents depthwise convolution with kernel size $k \times k$. The operator $[\cdot, \cdot]$ denotes channel-wise concatenation.

To model long-range interactions in the spectral domain, the aggregated local representation is transformed into Fourier space and processed by a learnable spectral response function:

$$\mathrm{X}_f = \mathcal{F}^{-1}(\Phi(\mathcal{F}(\mathrm{X}_l))) + \mathrm{X}_l \tag{2}$$

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the orthonormal real-valued Fourier transform and inverse transform, respectively, and $\Phi(\cdot)$ represents nonlinear spectral feature transformation. The residual connection preserves low-frequency structural consistency and stabilizes spectral reconstruction.

Subsequently, spatial–frequency coupled aggregation is performed via convolutional refinement:

$$\mathrm{Z} = \mathcal{H}\left(\mathrm{X}_f\right) \tag{3}$$

where $\mathcal{H}(\cdot)$ denotes local convolutional feature aggregation.

To further enhance informative feature responses, channel-adaptive recalibration is introduced:

$$\mathrm{A} = \sigma(\mathrm{W}_2\delta(\mathrm{W}_1\mathrm{GAP}(\mathrm{Z}))) \tag{4}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, $\delta(\cdot)$ denotes nonlinear activation, and $\sigma(\cdot)$ represents the sigmoid function.

Finally, the output of FFCM can be formulated as

$$\mathrm{Y} = \mathrm{A} \odot \mathrm{Z} \tag{5}$$

where $\odot$ denotes element-wise multiplication.

Compared with purely spatial convolutional operations, frequency-domain representations are less sensitive to illumination fluctuations and local texture degradation. Consequently, the proposed CSP-FFCM module enables the backbone network to better preserve discriminative contour and structural information of underground personnel under challenging conditions such as dust interference, low visibility, and pose variations. Furthermore, the integration of spatial and frequency representations improves feature diversity and enhances the robustness of subsequent detection stages.

4.2. POTE Module

The conventional RT-DETR model employs an Attention Information Fusion Interaction (AIFI) module, which excels in open-field scenarios but has notable limitations in underground mine environments. Detecting personnel around an LHD in a mine presents unique challenges, including non-uniform lighting, confined spaces, and complex backgrounds. The AIFI module, based on a standard multi-head self-attention mechanism, has a quadratic increase in computational complexity, represented as $O(N^2)$. More importantly, it struggles to effectively capture the local spatial and directional features that are crucial in these confined, complex settings. Furthermore, the traditional AIFI module is insufficient for modeling low-contrast targets, partially occluded personnel, and dynamic lighting changes that are common in underground scenes.

POLAR-DETR is a real-time end-to-end detection framework based on RT-DETR, designed for dense occlusion, small objects, and complex scenes in Total Laboratory Automation (TLA) environments. By introducing polarized occlusion-aware hierarchical feature encoding, multi-level hypergraph attention fusion, and Hessian-based pruning, it achieves 70.0% AP at 68.4 FPS on the AMPL dataset, offering an efficient solution for occluded object detection in industrial settings [38].

Personnel detection in underground environments requires a more precise perception of directional features and local spatial relationships while maintaining computational efficiency. Based on these considerations, we propose an improved module called POTE to mitigate the shortcomings of the conventional AIFI module in RT-DETR, thereby enhancing personnel detection performance around underground LHD. The proposed POTE module is adapted from the Polarity-aware Linear Attention mechanism introduced in PolaFormer [39]. In this work, the module is integrated into the RT-DETR architecture to replace the original Attention-based Intra-scale Feature Interaction (AIFI) module, thereby improving computational efficiency and directional feature modeling capability for underground personnel detection. As shown in Figure 12, the proposed POTE module primarily consists of a Polarity-aware Linear Attention and a Feed Forward network.

Polarity-aware Linear Attention is the core of the POTE module, designed to enhance the discriminative ability of feature selection while maintaining linear complexity. This transformation better aligns with the spatial distribution of image features, which is particularly effective for detecting underground personnel by more accurately measuring the correlations between different feature points. Next, linear transformations are used to compress and map the transformed features. This reduces computational redundancy and accelerates the generation of attention scores while retaining critical semantic and spatial information. This process provides higher-quality attention weight guidance for feature fusion.

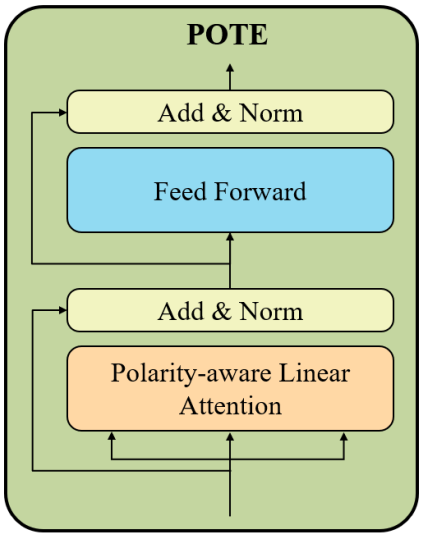


Figure 12. POTE Architecture Diagram.

As shown in Figure 13, unlike the standard self-attention mechanism, the Polarity-aware Linear Attention mechanism enhances its non-linear representation capability by decomposing the Query and Key vectors into their respective positive and negative polarities. This mechanism can effectively capture direction-sensitive feature representations while reducing computational complexity.

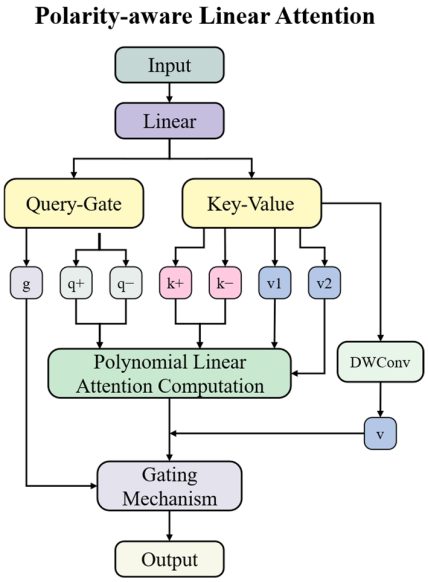


Figure 13. Detailed Diagram of Polarity-aware Linear Attention.

The mathematical expression for Polarity-aware Linear Attention is as follows: Given an input feature sequence $X \in \mathbb{R}^{N \times C}$, the polarity-aware linear attention first projects the input features into query, key, value, and gating branches:

$$Q, G = XW_{qg}, K, V = XW_{kv} \tag{6}$$

The input feature sequence is first linearly projected into query, key, value, and gating representations. The gating branch is used for adaptive feature modulation in the final stage.

To explicitly model heterogeneous semantic responses, polarity-aware decomposition is introduced:

$$Q^+ = \phi(Q)^p, Q^- = \phi(-Q)^p, K^+ = \phi(K)^p, K^- = \phi(-K)^p \tag{7}$$

where $\phi(\cdot)$ denotes the ReLU activation function. The dynamic exponent parameter is defined as $p = 1 + \alpha \cdot \sigma(P)$, where P is a learnable parameter and α controls the modulation strength.

Subsequently, two complementary query representations, namely similarity-aware and opposition-aware branches, are constructed to capture both homogeneous and heterogeneous semantic interactions.

$$Q_{\text{sin}} = [Q^+, Q^-], Q_{\text{opp}} = [Q^-, Q^+] \tag{8}$$

Instead of explicitly computing pairwise token correlations, the proposed method reformulates the attention operation into associative matrix multiplications, thereby reducing the computational complexity from quadratic to linear complexity with respect to the token number.

$$\begin{aligned} X_{sim} &= \frac{Q_{sim}\left(K_p^T V_1\right)}{Q_{sim}\overline{K}_p + \varepsilon} \\ X_{opp} &= \frac{Q_{opp}\left(K_p^T V_2\right)}{Q_{opp}\overline{K}_p + \varepsilon} \end{aligned} \tag{9}$$

Finally, the similarity-aware and opposition-aware attention features are concatenated and fused with the residual value branch. Finally, an adaptive gating mechanism is employed to enhance informative responses while suppressing irrelevant activations.

$$Y = \left([X_{sim}, X_{opp}] + V\right) \odot G \tag{10}$$

The POTE module proposed in this paper offers several advantages over the conventional AIFI module for detecting personnel in underground environments. First, its polarity-aware linear attention mechanism provides a direction-sensitive feature representation, which enhances the recognition of human posture and orientation in the complex underground setting. Compared with conventional self-attention, the proposed formulation avoids explicit pairwise token interaction computation and reduces the computational complexity from $O(N^2)$ to $O(N)$. The module also demonstrates enhanced local spatial modeling by combining depthwise separable convolution with the attention mechanism, which is crucial for detecting partially occluded personnel. Lastly, the introduction of a dynamic power parameter enables dynamic feature enhancement, allowing the model to adapt the intensity of feature transformations based on the input and, thereby, improving the overall robustness of personnel detection under varying and challenging lighting conditions.

4.3. MAL Function

In the task of underground personnel detection, the design of the loss function has a crucial impact on detection performance. The original RT-DETR framework employs Varifocal Loss (VFL), which, by coupling prediction confidence with true quality (e.g., IoU) in a weighted manner, enhances the contribution of high-quality samples to a certain extent. However, in complex underground scenarios, factors such as insufficient lighting, dust interference, and target occlusion often result in many low-quality negative samples. This can easily lead to a mismatch between positive and negative samples, thereby reducing detection accuracy and increasing the FPR.

In underground environments, the large number of easy negative samples often causes unstable classification confidence and increases the FPR. Although Varifocal Loss (VFL) emphasizes high-quality positive samples, it tends to suppress low-quality yet informative samples, resulting in a mismatch between classification confidence and localization quality. To address this issue, this paper replaces the VFL function with MAL, which effectively mitigates the problem of positive-negative sample mismatch [40]. Compared with VFL, MAL removes the balancing hyperparameter α and adopts a unified weighting strategy for both foreground and background samples, thereby simplifying the optimization process and improving training stability under severe foreground–background imbalance. The VFL can be formulated as:

$$\mathcal{L}_{\mathrm{VFL}}(p,q) = \begin{cases} -q[\mathrm{qlog}(p) + (1-q)\mathrm{log}(1-p)], & q > 0 \\ -\alpha p^{\gamma}\log(1-p), & q = 0 \end{cases} \tag{11}$$

where p denotes the predicted classification confidence, q represents the IoU score between the predicted bounding box and the corresponding ground-truth box, α is the balancing factor, and γ is the focusing parameter. For foreground samples $(q > 0)$, the target label is assigned as the IoU score q, while for background samples $(q = 0)$, the target label is set to zero.

To further enhance the consistency between classification confidence and localization quality, MAL introduces an alignment-aware scoring mechanism by replacing the target label q with $q\gamma$. This strategy reduces the excessive emphasis on high-quality samples while enhancing the contribution of low-quality but informative samples. Accordingly, the MAL is defined as:

$$\mathcal{L}_{\mathrm{MAL}}(p,q,y) = \begin{cases} -q^{\gamma}\log(p) - (1-q^{\gamma})\log(1-p), & y = 1 \\ -p^{\gamma}\log(1-p), & y = 0 \end{cases} \tag{12}$$

where $y \in 0, 1$ denotes the sample category indicator. Compared with VFL, MAL removes the balancing hyperparameter α and adopts a unified weighting strategy for positive and negative samples, thereby simplifying the optimization process.

In this study, the focusing parameter γ of MAL is set to 1.5, and the IoU threshold used for positive sample assignment is set to 0.7. These hyperparameter settings were selected to improve the stability of confidence estimation while maintaining sufficient sensitivity to difficult underground personnel samples. In summary, the proposed MAL effectively alleviates the positive-negative sample mismatch problem in underground personnel detection tasks.

5. Experiments and Results

5.1. Experimental Environment

To validate the performance of the proposed FP-DETR framework for underground personnel detection, extensive experiments were conducted under a unified hardware and software environment. The specific experimental system environment was a Windows 11 machine with an Intel(R) Core (TM) i9-14900KF CPU, an NVIDIA GeForce RTX5080 GPU, and 16 GB of RAM. The model was trained using Python 3.9, CUDA 12.8, and PyTorch 2.7. For detailed information on the hyperparameter settings, please refer to Table 3. These parameters were based on the configuration of the original RT-DETR model and determined through empirical experimentation. For model evaluation, the confidence threshold was set to 0.25 during inference. FP-DETR and RT-DETR were evaluated without any additional NMS post-processing, preserving the end-to-end nature of DETR-based detection. All experimental results were evaluated using the same inference settings to

ensure fair comparison among different methods. The proposed method was trained and evaluated using the underground personnel dataset described in Section 3.

Table 3. Model hyperparameter settings.

Configurations	Values
Image size	640×640
Epochs	150
Batch Size	8
Optimizer	AdamW
Learning Rate	1×10^{-4}
Weight Decay	5×10^{-4}

5.2. Evaluation Metrics

This paper utilized a variety of evaluation metrics to assess the improved model, including the number of parameters (Params), giga floating-point operations (GFLOPs), and average precision (AP). These metrics are widely used in the field of object detection to evaluate both model accuracy and efficiency.

Params reflect the storage requirements of the model, while GFLOPs is used to quantify its computational complexity. AP is widely considered a key indicator of performance for object detection tasks. AP is measured by calculating the area under the precision-recall curve, which provides insight into the overall detection accuracy of the model. The formula for calculating AP is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$AP = \int_0^1 Precision(r) dRecall(r) \tag{15}$$

In this paper, the model's performance was evaluated using the mAP metric, specifically at two different Intersection over Union (IoU) thresholds: mAP@0.5 and mAP@0.5:0.95. The mAP@0.5 metric provides a measure of the mAP at a fixed IoU threshold of 0.5 and is commonly utilized for a rapid assessment of performance. For a more comprehensive evaluation, the mAP@0.5:0.95 metric was employed. This metric calculates the average mAP across ten distinct IoU thresholds, ranging from 0.5 to 0.95 in increments of 0.05. This approach facilitates a more robust evaluation of the model's accuracy across varying degrees of detection stringency. As the dataset utilized in this research contains only a single class (miners), the resulting mAP and AP values are considered identical.

5.3. Results and Analysis

To validate the efficacy of our proposed algorithm, we conducted five experiments aimed at assessing its performance in terms of detection accuracy and false positives. These experiments are structured as follows: (1) Comparison among enhancement schemes of different backbone networks; (2) Comparative analysis of various attention information fusion and interaction modules; (3) Ablation studies to dissect the specific contributions of each module and their synergistic effects; (4) Benchmarking the detection accuracy of the proposed algorithm against mainstream approaches; and (5) Evaluation of the algorithm's false alarm rate and robustness in genuine underground environments. These comprehensive experiments were designed to thoroughly assess the performance of our

approach in complex scenarios, thereby substantiating the advantages introduced by our improvements.

5.3.1. Comparison Using Different Backbone Networks

Figure 14 illustrates the evolution of mAP during training when RT-DETR is enhanced with different backbone networks—namely, GlobalFilter, FasterFDConv, ConvNeXt, and the proposed FFCM. As shown, all variants converge after approximately 120 training epochs. Among them, the model equipped with our proposed FFCM module achieves the highest performance in both mAP@0.5 and mAP@0.5:0.95. Notably, the margin of improvement is particularly pronounced in mAP@0.5:0.95, outperforming the other backbone variants. This result demonstrates that FFCM possesses better feature extraction capability under the challenging conditions of underground environments, where complex lighting, occlusions, and texture variations are prevalent.

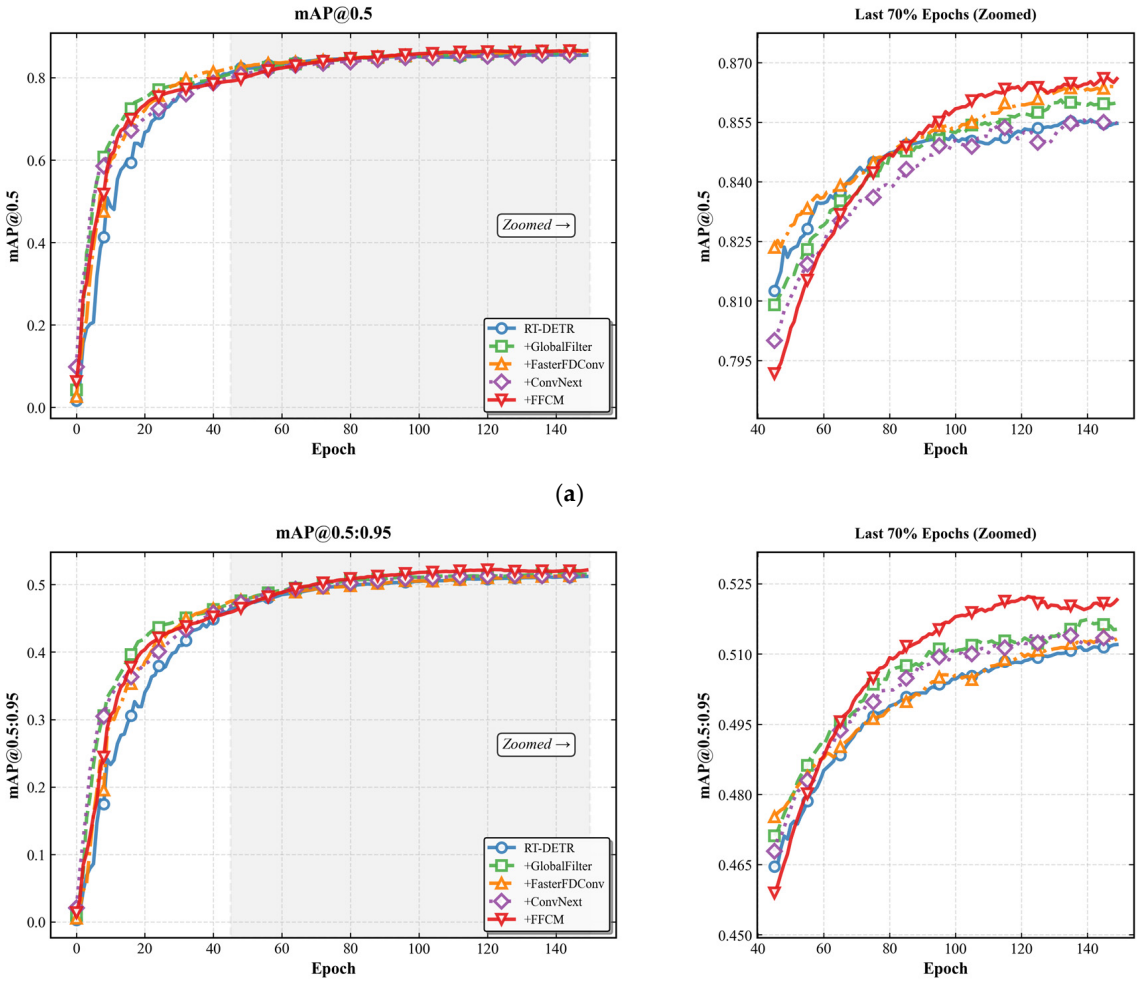


Figure 14. Evolution of mAP metrics on the validation set for different backbone networks over training epochs. (**a**) mAP@0.5 of different backbone networks on the validation set; (**b**) mAP@0.5:0.95 of different backbone networks on the validation set. The grey shaded region marks the area enlarged in the right panel, and the arrows indicate the corresponding detailed view shown on the right.

Figure 15 shows the precision–recall curves of the four RT-DETR variants, all of which stabilize after approximately 100 training epochs with no significant post-convergence divergence. Among them, ConvNeXt exhibits higher precision fluctuations and consistently lower accuracy than both the baseline and other variants. The proposed FFCM achieves

precision comparable to FasterFDConv—among the highest observed—while its recall is only slightly lower than that of GlobalFilter, indicating a favorable balance between minimizing false positives and maintaining detection coverage, particularly advantageous in safety-sensitive underground environments.

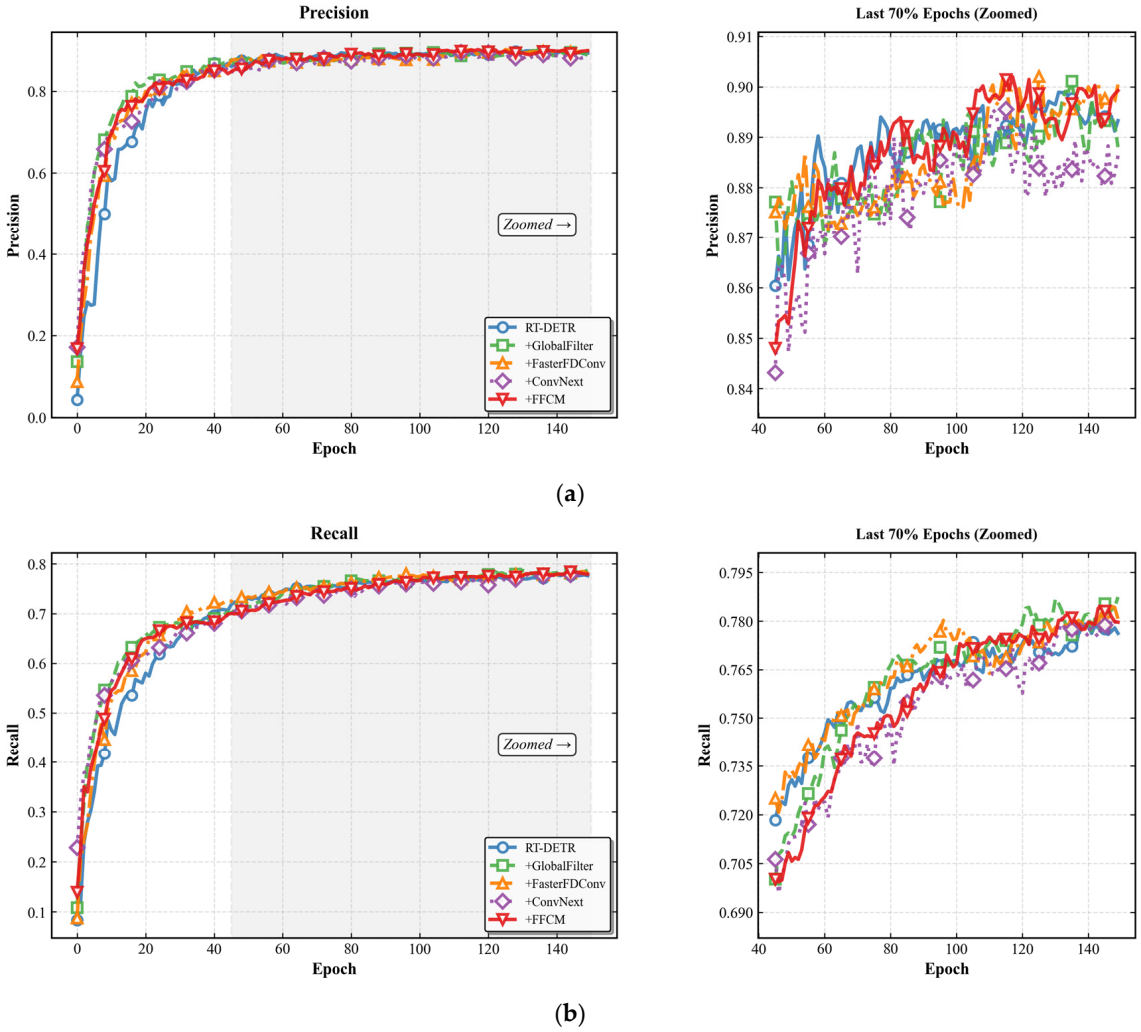


Figure 15. Evolution of precision (P) and recall (R) on the validation set for different backbone networks over training epochs. (**a**) Precision of different backbone networks on the validation set; (**b**) Recall of different backbone networks on the validation set. The grey shaded region marks the area enlarged in the right panel, and the arrows indicate the corresponding detailed view shown on the right.

As shown in Table 4, the proposed FFCM module achieves the highest mAP@0.5 (0.865) and mAP@0.5:0.95 (0.521) among all variants while maintaining a comparable model size (15.6 M parameters) and computational cost (50.4 GFLOPs). Although GlobalFilter and FasterFDConv offer marginal gains in mAP@0.5, they lag behind FFCM—particularly in the more rigorous mAP@0.5:0.95 metric. The improved performance of FFCM is attributed to its Fourier-based feature representation, which enhances robustness to low-light, noise, and texture ambiguity in underground environments. This demonstrates the performance of frequency-domain modeling for complex-scene object detection, leading to FFCM's selection as the backbone in our final framework.

Table 4. Comparison of detection results using different backbone networks.

Method	Params (M)	GFLOPs	mAP0.5	mAP0.5:0.95	P	R
RT-DETR	20.0	57.3	0.857	0.515	0.886	0.785
+GlobalFilte	12.4	33.2	0.861	0.517	0.910	0.787
+FasterFDConv	13.3	44.2	0.864	0.513	0.911	0.784
+ConvNeXt	16.9	46.8	0.859	0.515	0.884	0.777
+FFCM	15.6	50.4	0.865	0.521	0.899	0.779

5.3.2. Comparison Among Different Attention Modules

Figure 16 illustrates the evolution of mAP during training when RT-DETR is enhanced with different attention-based feature fusion and interaction modules—namely, DHSA, DPB, CAB, and the proposed POTE. All variants exhibit convergence behavior after approximately 80 training epochs. In terms of mAP@0.5, all four modules yield comparable improvements over the baseline, with marginal performance differences among them. However, under the more stringent mAP@0.5:0.95 metric, notable distinctions emerge: DHSA and CAB achieve similar levels of performance, while the proposed POTE module consistently attains the highest mAP@0.5:0.95, demonstrating its improved capability in enabling precise object localization across a range of IoU thresholds. This highlights the performance of the POTE design in facilitating more discriminative and robust feature interactions for accurate detection in complex scenarios.

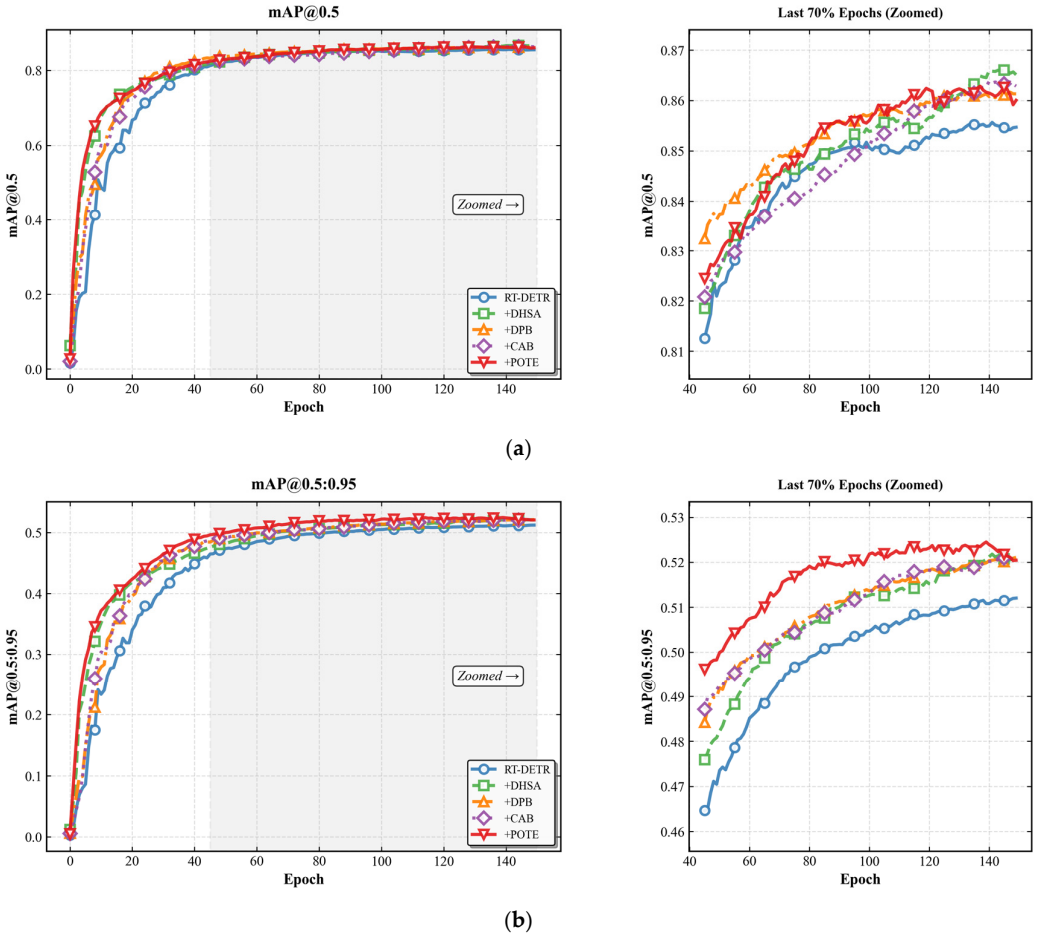


Figure 16. Evolution of mAP metrics on the validation set for different attention modules over training epochs. (**a**) mAP@0.5 of different attention modules on the validation set; (**b**) mAP@0.5:0.95 of different attention modules on the validation set. The grey shaded region marks the area enlarged in the right panel, and the arrows indicate the corresponding detailed view shown on the right.

Figure 17 illustrates the precision–recall curves of RT-DETR enhanced with different attention modules (DHSA, DPB, CAB, and the proposed POTE). All methods stabilize after approximately 100 epochs with no significant post-convergence trend differences. DHSA and DPB exhibit larger precision fluctuations and lower accuracy compared to POTE. Although DHSA achieves a higher recall peak around 130 epochs, it subsequently declines, whereas POTE shows a steadier recall performance with minor fluctuations and an overall upward trend before stabilizing. This indicates that POTE not only improves precision but also ensures more reliable and stable recall, making it well-suited for complex detection tasks.

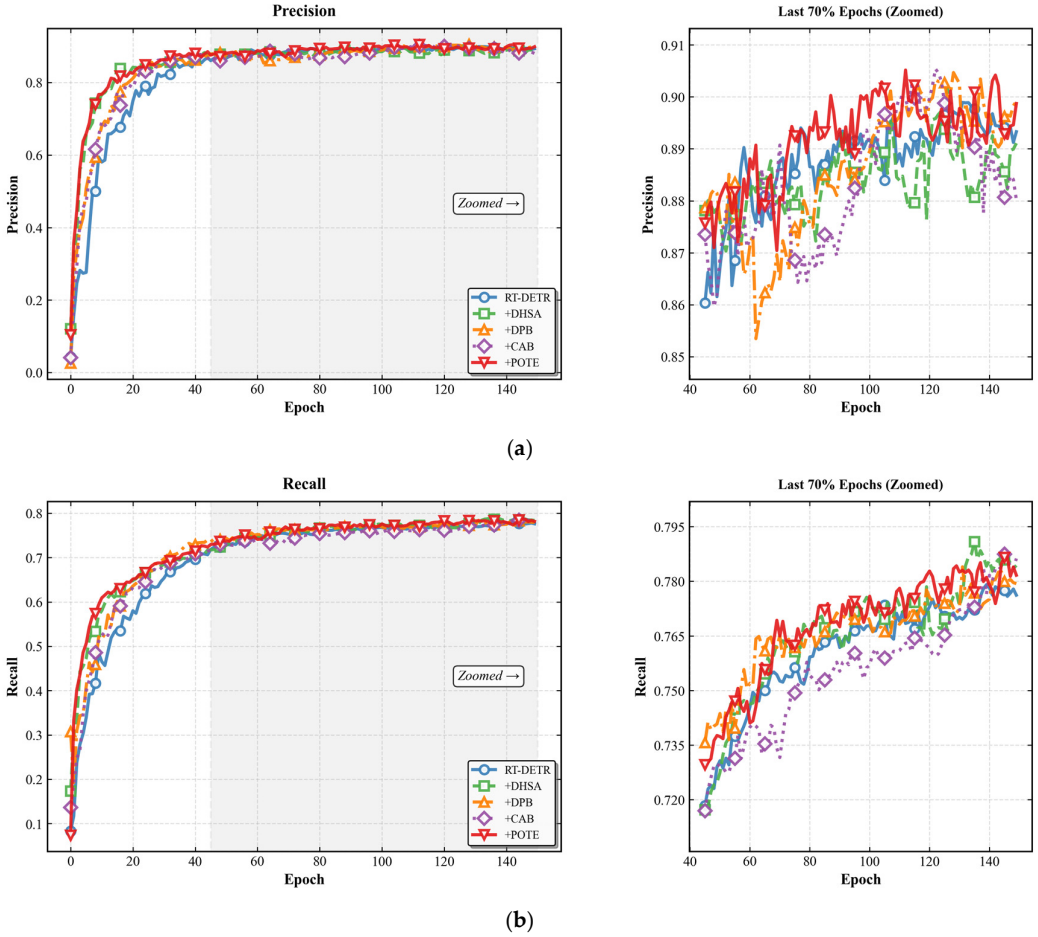


Figure 17. Evolution of precision (P) and recall (R) on the validation set for different attention modules over training epochs. (**a**) Precision of different attention modules on the validation set; (**b**) Recall of different attention modules on the validation set. The grey shaded region marks the area enlarged in the right panel, and the arrows indicate the corresponding detailed view shown on the right.

As shown in Table 5, while DHSA, DPB, and CAB yield modest improvements in mAP@0.5 over the baseline RT-DETR, the proposed POTE module achieves the highest performance on the more comprehensive metric mAP@0.5:0.95 (0.524), representing a +0.9 percentage point gain relative to the baseline (0.515)—the largest among all variants—without incurring additional computational overhead (20.1 M parameters, 57.6 GFLOPs). This indicates that POTE enhances detection robustness across a range of IoU thresholds by more effectively integrating intra- and inter-layer features, thereby preserving both precise localization cues and rich semantic information. Despite comparable model complexity, POTE consistently demonstrates better performance in mAP@0.5:0.95, highlighting its

architectural efficiency and ability to enhance feature interaction fidelity in complex real-world scenarios.

Table 5. Comparison of detection results using different attention interaction modules.

Method	Params (M)	GFLOPs	mAP0.5	mAP0.5:0.95	P	R
RT-DETR	20.0	57.3	0.857	0.515	0.886	0.785
+DHSA	20.1	57.6	0.865	0.520	0.883	0.791
+DPB	19.9	57.5	0.861	0.521	0.900	0.784
+CAB	20.1	61.4	0.863	0.520	0.902	0.782
+POTE	20.1	57.6	0.862	0.524	0.898	0.781

5.3.3. Ablation Experiments

To investigate the performance and contribution of each innovative module in the proposed FP-DETR algorithm for detecting personnel around underground LHDs, we conducted a series of ablation experiments. This was done by comparing the performance of different model configurations across various evaluation metrics. Specifically, we assessed the impact of the CSP-FFCM module, the improved POTE module, and the introduction of the MAL function on the model's detection performance and further analyzed the synergistic effect when all three modules were combined.

The experiments were conducted on a unified training dataset composed of the RTTS dataset and a self-collected dataset to ensure consistency across all model variants. All models were trained under identical conditions, including the same training strategy, hyperparameter settings, and hardware environment, to guarantee fairness and comparability. To assess the stability and reproducibility of the proposed method, each experiment was repeated three times using different random seeds (666, 777, and 888), and the mean and standard deviation of all evaluation metrics were reported. The evaluation metrics included accuracy, recall, mAP at IoU thresholds of 0.5 and 0.5:0.95, GFLOPs, and the number of parameters (Params).

To comprehensively evaluate the robustness and domain adaptability of the proposed FP-DETR, three validation settings were designed: (1) validation on the self-collected dataset only, (2) validation on the RTTS-only dataset, and (3) validation on a merged-domain dataset combining both RTTS and self-collected samples. These settings enable a systematic assessment of model performance under single-domain and cross-domain conditions.

In our ablation experiments, we constructed the following model variants for comparison:

Model 1 (Baseline): The original RT-DETR-r18 model without any of the proposed improvements.

Model 2 (+FFCM): The Baseline model with the FFCM module incorporated to build a new backbone, replacing the original ResNet-18 backbone.

Model 3 (+POTE): The Baseline model with the improved POTE module replacing the multi-head self-attention mechanism in the original AIFI module.

Model 4 (+MAL): The Baseline model with the MAL function replacing the original VFL.

Model 5 (+FFCM + POTE + MAL): The complete FP-DETR model, which includes the CSP-FFCM module, the improved POTE module, and the MAL function.

The experimental results, shown in Table 6, demonstrate that each modification contributes differently to improving the model's detection performance.

Table 6. Performance Comparison of Different Model Variants on underground-only validation set.

Model	CSP-FFCM	POTE	MAL	Params (M)	GFLOPs	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%
Baseline	-	-	-	20.0	57.3	90.1 (±0.14)	84.3 (±0.11)	87.0 (±0.17)	55.1 (±0.08)
Baseline + FFCM	√	-	-	15.6	50.4	90.2 (±0.12)	86.4 (±0.09)	87.8 (±0.21)	55.8 (±0.16)
Baseline + POTE	-	√	-	20.1	57.6	91.3 (±0.11)	88.4 (±0.18)	88.3 (±0.15)	57.7 (±0.11)
Baseline + MAL	-	-	√	20.0	57.3	90.4 (±0.08)	87.2 (±0.13)	87.7 (±0.24)	56.3 (±0.12)
Baseline +FFCM +POTE +MAL	√	√	√	15.8	51.2	91.5 (±0.12)	88.7 (±0.10)	88.5 (±0.18)	58.6 (±0.15)

Note: "-" indicates that the corresponding module is not used, while "√" indicates that the corresponding module is included in the model variant.

The experimental results presented in Table 6 demonstrate that each proposed component contributes differently to improving the detection performance of RT-DETR for underground personnel detection on the underground-only validation set.

The CSP-FFCM module improves the feature extraction capability of the backbone network while significantly reducing model complexity. Compared with the baseline RT-DETR model, the introduction of CSP-FFCM increases the mAP@0.5 from 87.0% (±0.17) to 87.8% (±0.21), while reducing the number of parameters from 20.0 M to 15.6 M and decreasing GFLOPs from 57.3 to 50.4. In addition, the recall improves from 84.3% (±0.11) to 86.4% (±0.09), demonstrating that the reconstructed backbone enhances feature extraction capability for underground personnel while maintaining lightweight characteristics. The relatively small standard deviation further indicates that the reconstructed backbone provides stable detection performance under different random initializations.

After replacing the original AIFI module with the POTE module, the model achieves consistent improvements in both mAP@0.5 and mAP@0.5:0.95, reaching 88.3% (±0.15) and 57.7% (±0.11), respectively. Meanwhile, precision and recall increase to 91.3% (±0.11) and 88.4% (±0.18), respectively. These results demonstrate that the polarity-aware linear attention mechanism improves global feature interaction and enhances feature representation capability for underground personnel under low-illumination and occlusion conditions.

When the MAL function is introduced, the model achieves an mAP@0.5 of 87.7% (±0.24) and an mAP@0.5:0.95 of 56.3% (±0.12), showing improved confidence–quality alignment and better optimization stability compared with the baseline model. In addition, the recall increases from 84.3% (±0.11) to 87.2% (±0.13), indicating that the MAL effectively alleviates the imbalance between foreground and background samples in underground environments. Since the MAL module does not introduce additional parameters or computational complexity, it improves detection performance without increasing model burden.

By integrating CSP-FFCM, POTE, and MAL simultaneously, the proposed FP-DETR achieves the best overall performance, reaching 88.5% (±0.18) mAP@0.5 and 58.6% (±0.15) mAP@0.5:0.95. Meanwhile, precision and recall further improve to 91.5% (±0.12) and 88.7% (±0.10), respectively. Notably, the parameter count and computational complexity remain lower than those of the baseline RT-DETR-r18 model, with only 15.8 M parameters and 51.2 GFLOPs. The relatively small experimental variance demonstrates that the proposed framework achieves stable and consistent performance improvements across different training runs.

The experimental results presented in Table 7 further evaluate the robustness and cross-domain generalization capability of the proposed FP-DETR on the RTTS-only dataset and the merged-domain validation set.

Table 7. Performance Comparison of Different Model Variants on RTTS-only dataset and merged-domain validation set.

Model	RTTS-Only Dataset		Merged-Domain Validation Set	
	mAP@ 0.5/%	mAP @0.5:0.95/%	mAP@ 0.5/%	mAP @0.5:0.95/%
Baseline	80.8 (±0.12)	50.1 (±0.10)	85.7 (±0.15)	51.5 (±0.11)
Baseline + FFCM	80.3 (±0.10)	51.4 (±0.17)	86.5 (±0.11)	52.1 (±0.13)
Baseline + POTE	81.8 (±0.08)	51.7 (±0.15)	86.2 (±0.13)	52.4 (±0.10)
Baseline + MAL	81.1 (±0.15)	51.4 (±0.08)	86.7 (±0.17)	51.7 (±0.12)
Baseline +FFCM +POTE +MAL	82.4 (±0.14)	52.4 (±0.12)	87.3 (±0.22)	52.5 (±0.12)

On the RTTS-only dataset, the proposed FP-DETR achieves the best overall performance, reaching 82.4% (±0.14) mAP@0.5 and 52.4% (±0.12) mAP@0.5:0.95, outperforming the baseline RT-DETR model by 1.6% and 2.3%, respectively. Although RTTS is not an underground mining dataset, it contains severe haze and degraded visibility conditions, which can effectively evaluate the robustness of the proposed method under adverse visual environments. The experimental results demonstrate that the POTE module contributes significantly to improving feature interaction and contextual representation under low-visibility conditions, while the MAL further enhances optimization stability and detection robustness. Although the CSP-FFCM module slightly decreases mAP@0.5 on the RTTS-only dataset, it improves mAP@0.5:0.95 from 50.1% (±0.10) to 51.4% (±0.17), indicating improved localization capability under complex visibility conditions.

On the merged-domain validation set, all proposed modules consistently improve detection performance compared with the baseline model. The complete FP-DETR achieves 87.3% (±0.22) mAP@0.5 and 52.5% (±0.12) mAP@0.5:0.95, demonstrating improved adaptability when simultaneously facing underground scenes and real-world hazy environments. In particular, the MAL module improves confidence–quality alignment, while the POTE module effectively enhances global feature interaction across different visual domains. These results demonstrate that the proposed FP-DETR maintains stable detection performance under diverse and cross-domain scenarios.

By integrating these three components, the proposed FP-DETR consistently achieves the best overall performance across different validation settings, confirming the complementary and synergistic advantages of the proposed improvements for underground personnel detection in complex environments.

Figure 18 compares the performance of five model variants—RT-DETR, +FFCM, +POTE, +MAL, and FP-DETR—showing that the full model achieves the highest precision, recall, and mAP while significantly reducing parameters. The FFCM backbone improves recall and efficiency, the POTE module notably boosts mAP, and the MAL further enhances detection accuracy. Ablation studies confirm that these three components provide complementary and synergistic improvements, underscoring their essential roles in optimizing RT-DETR for complex detection scenarios.

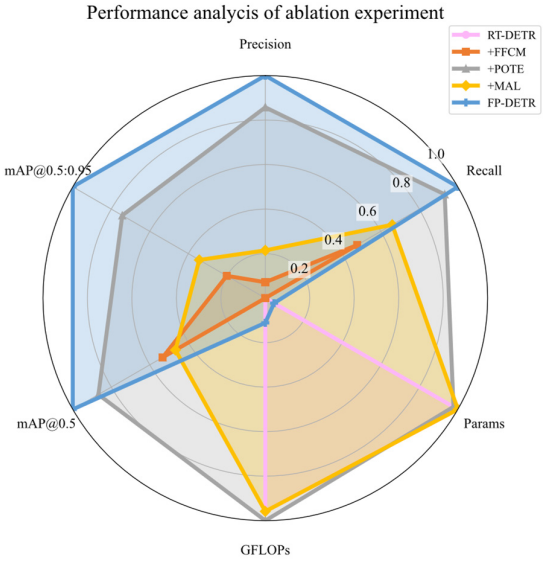


Figure 18. Comparison of Model Performance and Parameter Count with Different Modules.

As illustrated in Figure 19, the normalized bar chart further compares the performance of different model variants in terms of Precision, Recall, Params, GFLOPs, mAP@0.5, and mAP@0.5:0.95. Compared with the baseline RT-DETR, the CSP-FFCM module improves recall while reducing model complexity, demonstrating its effectiveness in lightweight feature extraction. The POTE module significantly enhances mAP performance by improving global feature interaction and directional perception. Furthermore, the MAL improves precision and detection stability by alleviating the mismatch between classification confidence and localization quality. By integrating these three components, the proposed FP-DETR achieves the best overall balance between detection accuracy and computational efficiency, confirming the complementary advantages of each module.

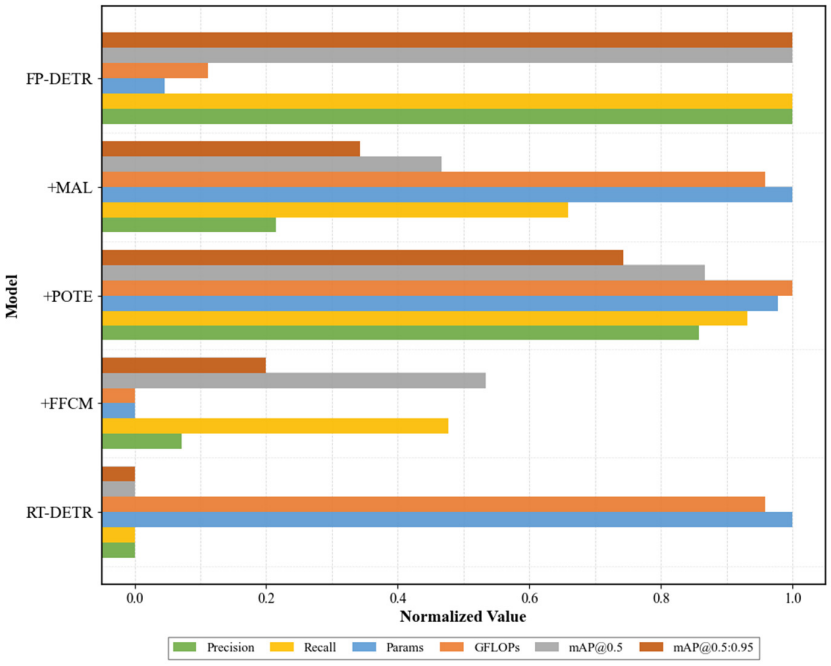


Figure 19. Bar chart comparing model performance and number of parameters after normalization for different modules.

In summary, the ablation experiments validate that the proposed FFCM module, the improved POTE module, and the MAL function each play a crucial role in improving

the model's performance. Their combined effect provides consistent evidence for these improvements in optimizing the RT-DETR model.

5.3.4. Effect of RTTS-Assisted Training

To further evaluate the performance of RTTS-assisted training for underground personnel detection, we conducted comparative experiments with and without the RTTS dataset under the same training settings. The experimental results are presented in Table 8.

Table 8. Effect of RTTS-assisted Training on Underground Detection Performance.

Model	With RTTS		Without RTTS	
	mAP@ 0.5/%	**mAP @0.5:0.95/%**	**mAP@ 0.5/%**	**mAP @0.5:0.95/%**
Baseline	87.0 (±0.17)	55.1 (±0.07)	85.8 (±0.24)	53.4 (±0.18)
Baseline + FFCM	87.8 (±0.14)	55.8 (±0.11)	86.3 (±0.16)	54.1 (±0.09)
Baseline + POTE	88.3 (±0.09)	57.7 (±0.18)	86.9 (±0.14)	55.6 (±0.11)
Baseline + MAL	87.7 (±0.12)	57.3 (±0.13)	86.5 (±0.15)	55.0 (±0.19)
Baseline +FFCM +POTE +MAL	88.5 (±0.16)	58.6 (±0.13)	87.2 (±0.11)	56.7 (±0.10)

The results show that incorporating the RTTS dataset consistently improves the detection performance of all model variants on the underground-only validation set. Compared with training without RTTS, the proposed FP-DETR achieves an improvement from 87.2% to 88.5% in mAP@0.5 and from 56.7% to 58.6% in mAP@0.5:0.95 after introducing RTTS-assisted training. These results indicate that the RTTS dataset provides beneficial degraded-visibility features and enhances the robustness and generalization capability of the proposed method under complex underground environments. Meanwhile, the proposed FP-DETR still maintains strong detection performance even without RTTS training, demonstrating that the performance improvements mainly originate from the proposed network design rather than reliance on external datasets.

5.3.5. Comparative Experiments

To validate the performance and superiority of the proposed FP-DETR algorithm for detecting personnel around underground LHDs, we meticulously designed a series of comparative experiments. We selected several advanced and distinct object detection models to compare their performance with our improved algorithm. All compared methods were trained under identical settings, including input resolution, training epochs, and data augmentation strategies. In addition, none of the models used pretrained weights, ensuring a fair comparison under the same training conditions. The specific models chosen for comparison are as follows.

The two-stage detection algorithm, Faster R-CNN, offers high accuracy, making it suitable for scenarios with extremely high precision requirements. However, its slow speed makes it unsuitable for the real-time demands of an underground environment. In contrast, the one-stage detection algorithm, TOOD, is fast and can promptly handle dynamically changing underground scenes, though its large number of model parameters and high computational load make it difficult to deploy. Additionally, the widely applied YOLOv8m, YOLO11m and YOLOX models from the YOLO series can complete detection in a short time, meeting the need for rapid detection of personnel around underground LHDs (Load–Haul–Dump machines), but their detection capability in complex underground environments still has room for improvement. The Transformer-based Deformable-DETR and the benchmark

model RT-DETR-r18 are effective at capturing global features, giving them an advantage when dealing with complex scenes, but their robustness in such scenarios needs further enhancement. To ensure fairness and comparability, all experiments were conducted using the same dataset and hardware environment. The comparative results are shown in Table 9.

Table 9. Performance Comparison on Underground-only Validation Set.

Method	Params (M)	GFLOPs	mAP@0.5/%	mAP@0.5:0.95/%
Faster R-CNN	41.4	207	86.3 (±0.08)	54.8 (±0.06)
TOOD	32.2	198	86.9 (±0.17)	56.1 (±0.13)
Deformable-DETR	40.1	195	86.7 (±0.22)	54.4 (±0.15)
RT-DETR-r18	20.0	57.3	87.0 (±0.17)	55.1 (±0.08)
YOLOX	5.0	7.57	85.1 (±0.19)	54.2 (±0.14)
YOLO11m	20.1	68	86.4 (±0.21)	56.0 (±0.11)
YOLOv8m	25.9	78.9	86.9 (±0.18)	56.7 (±0.14)
FP-DETR	15.8	51.2	88.5 (±0.20)	58.6 (±0.13)

The experimental results presented in Table 9 demonstrate that the proposed FP-DETR achieves the best overall detection performance among all compared methods on the underground-only validation set. Specifically, FP-DETR achieves an mAP@0.5 of 88.5% (±0.20) and an mAP@0.5:0.95 of 58.6% (±0.13), outperforming the baseline RT-DETR-r18 model, which achieves 87.0% (±0.17) and 55.1% (±0.08), respectively.

Compared with representative CNN-based detectors such as Faster R-CNN, TOOD, YOLOX, YOLO11m, and YOLOv8m, the proposed FP-DETR achieves improved detection accuracy while maintaining relatively low model complexity. FP-DETR surpasses YOLOv8m by 1.6% in mAP@0.5 and 1.9% in mAP@0.5:0.95, while requiring fewer parameters and lower computational cost. Compared with Transformer-based methods such as Deformable-DETR and RT-DETR-r18, the proposed method further improves localization accuracy and robustness under complex underground environments.

Meanwhile, the proposed FP-DETR requires only 15.8 M parameters and 51.2 GFLOPs, representing approximately a 21% reduction in parameters and lower computational complexity compared with the original RT-DETR-r18 model. Although lightweight detectors such as YOLOX exhibit lower computational complexity, their detection accuracy remains inferior to the proposed method.

Overall, the experimental results demonstrate that FP-DETR achieves a better balance between detection accuracy, computational efficiency, and model complexity, demonstrating promising potential for underground personnel detection, although broader field validation remains necessary.

As shown in Figure 20, the proposed FP-DETR model demonstrates better performance across multiple key metrics, outperforming state-of-the-art detection architectures such as YOLOX, YOLOv8m, YOLO11m, RT-DETR-r18, Deformable-DETR, TOOD, and Faster R-CNN. Specifically, the proposed method achieves the highest mAP@0.5 and mAP@0.5:0.95, indicating its reliable accuracy in object detection under both moderate and stringent IoU thresholds. This outstanding precision detection is attributed to the model's optimized architectural design, which moderately enhances feature representation and localization capability.

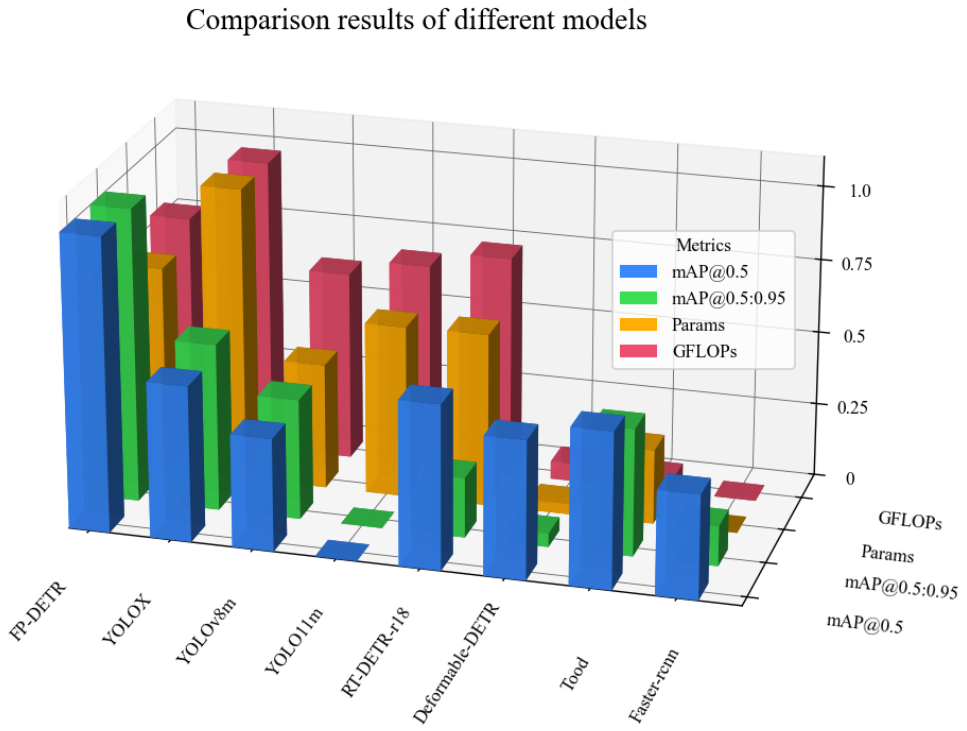


Figure 20. Normalized overall metrics compared with mainstream models.

In summary, these results suggest that FP-DETR is a promising approach for underground personnel detection. These results underscore the performance of the proposed architecture in balancing performance and scalability, positioning it as a promising advancement in real-time detection systems.

To visually evaluate the performance of our proposed FP-DETR model against the baseline RT-DETR and YOLO11m models, we conducted a comparative analysis using image detection visualizations. As shown in Figure 21, a clear contrast is presented, with each set of images displaying the original image, followed by the detection results from YOLO11m, RT-DETR, and FP-DETR, respectively., in bright scenes, while YOLO11m has no obvious false positives, its detection results are often inaccurate, with issues of duplicate detections and low confidence. The RT-DETR model, on the other hand, shows clear instances of false positive detections. Our improved model achieves higher detection confidence and does not produce false positives. In dusty environments with strong background interference and high image blurriness, which demand high algorithm robustness, both YOLO11m and RT-DETR-r18 exhibit missed detections. Our proposed model, however, does not miss any targets in these scenes, accurately separating them from interference and demonstrating strong ant-interference capabilities. For low-light conditions, where images are obscured by darkness and dominated by noise, YOLO11m and RT-DETR-r18 often suffer from duplicate detections and false positives due to insufficient feature extraction. Our model, by introducing the CSP-FFCM module, moderately enhances feature representation in crucial areas and shows good robustness in such low-light tests. Finally, in occlusion scenes, where targets are partially blocked and structural information is incomplete, YOLO11m and RT-DETR-r18 show a significant drop in accuracy and produce duplicate detections. Our proposed model, by incorporating a context completion mechanism and an extreme-point-aware attention fusion strategy, moderately recovers discriminative features in occluded regions. It maintains high detection accuracy, outperforming existing comparative models and demonstrating an excellent ability to model target integrity.

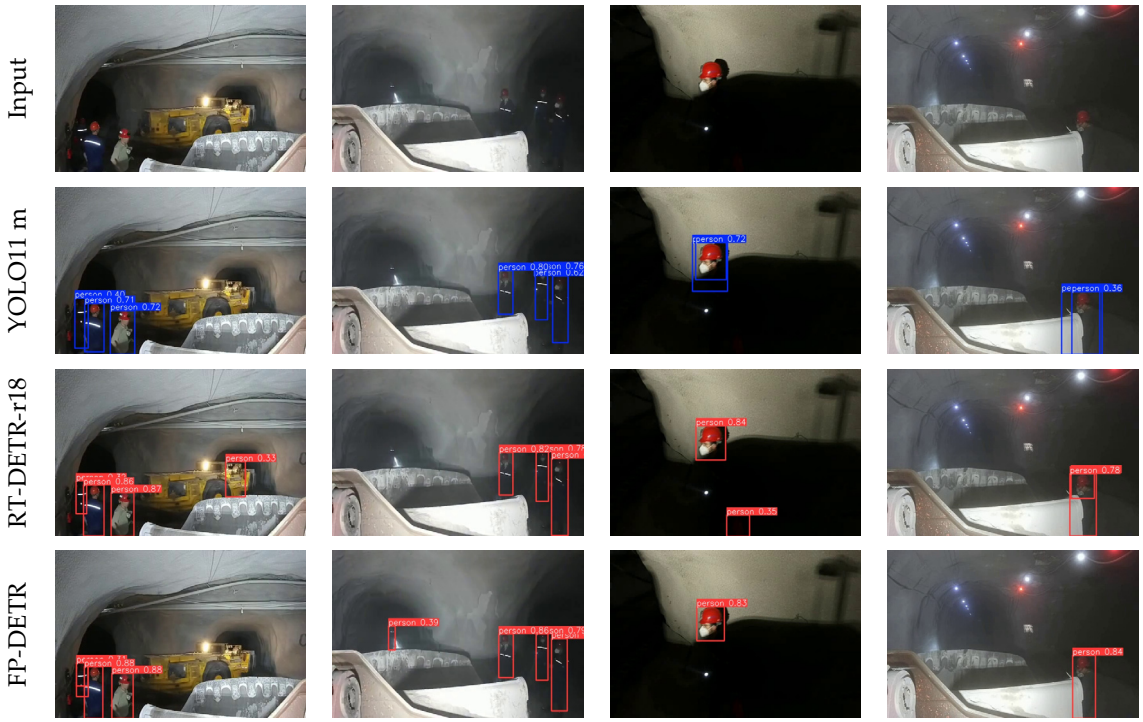


Figure 21. Visual Comparison of Detection Results for Different Models. (**a**) Bright scene; (**b**) Dusty scene; (**c**) Low-light scene; (**d**) Occluded scene.

As shown in Figure 22, a heatmap visualization analysis of different models' detection results was conducted to clearly demonstrate the learning capability of our improved model. The images reveal that YOLO11m and RT-DETR frequently learn incorrect information in underground environments, leading to missed and false detections. In contrast, our proposed method can accurately learn human-specific information, creating a much clearer distinction from the background and making its detection performance more suitable for underground settings. In bright scenes, our model accurately focuses on targets in corners, and the generated attention regions are more compact and defined, with a better response to target edges than the comparison models. In dusty scenes, where heavy visual blur and granular noise pose a challenge, our model's heatmap distribution is clearly concentrated on the target area and highly consistent with the actual boundaries, while the comparison models show missed detections in the noisy regions. For low-light conditions, where traditional methods often exhibit sparse heatmap distributions and unstable attention areas leading to false positives, our model maintains a high target response intensity and accurately locates human regions, demonstrating stronger illumination invariance and feature representation ability. In occlusion scenarios, where traditional methods often fail to produce a continuous response due to partial blocking, our method, by introducing a context modeling and region completion mechanism, maintains a high response probability in occluded areas. This allows it to successfully identify partially blocked human silhouettes, showcasing improved perceptual integrity. The combined results from these typical underground environments indicate that the proposed FP-DETR achieves more stable detection performance than the compared methods under complex scenarios, while maintaining relatively higher detection accuracy and a lower FPR. These results suggest that the proposed method has good adaptability and robustness for practical underground deployment conditions.

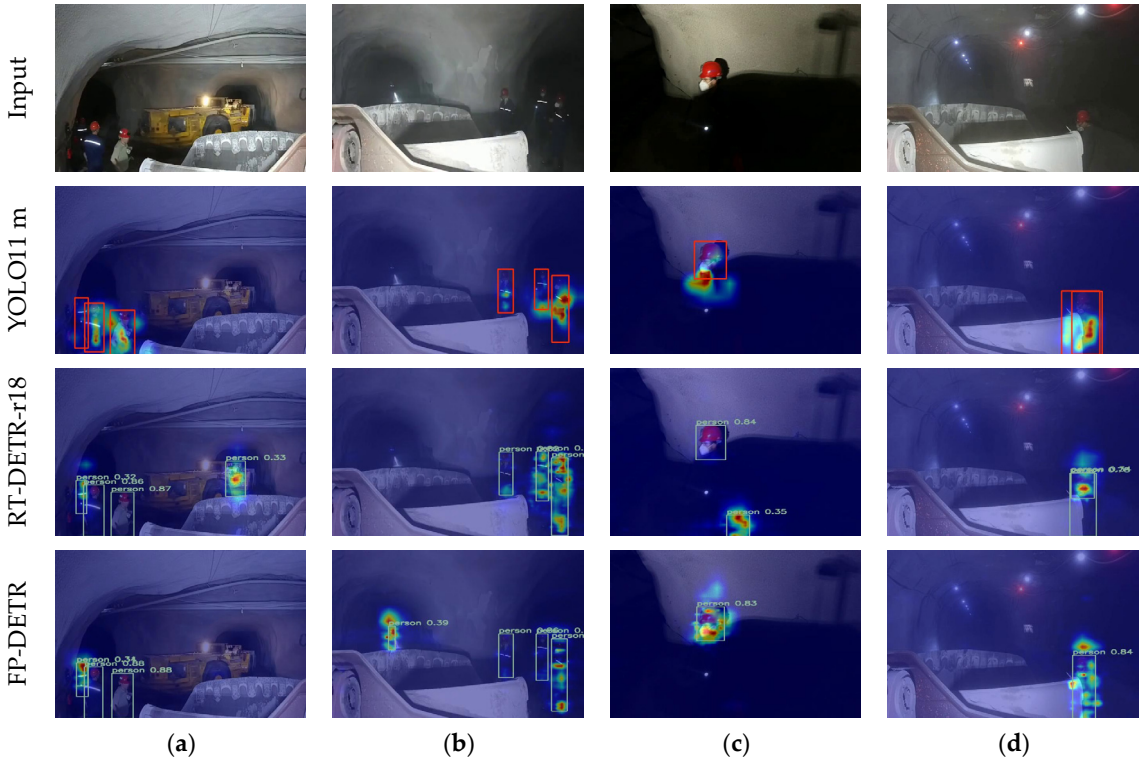


Figure 22. Comparison of Detection Heatmaps for Different Models. (**a**) Bright scene; (**b**) Dusty scene; (**c**) Low-light scene; (**d**) Occluded scene. The red boxes in the image indicate the people detected.

5.3.6. Field Experiment

The comparison experiments presented in the previous section demonstrate that the proposed FP-DETR improves detection accuracy while reducing computational complexity. To further evaluate false alarm behavior in practical underground environments, a one-day field experiment was conducted at the Magang Luohe Iron Mine in Hefei City, Anhui Province, using both the baseline RT-DETR-r18 and the proposed FP-DETR models. A main tunnel along a typical LHD operating route was selected for testing, and the models were evaluated using the frame-level FPR.

For quantitative evaluation experiments in previous sections, the confidence threshold was set to 0.25 to maintain consistency with standard model comparison settings. In contrast, for field experiments and visualization analysis, the confidence threshold was set to 0.3 to suppress unstable low-confidence detections and better reflect the practical deployment requirements of underground safety monitoring systems. All models used the same confidence threshold during field testing to ensure fair comparison.

In this study, FPR is defined as the proportion of frames incorrectly predicted as containing personnel among all frames without human presence. Specifically, the same video frames were processed by different models, and the number of false positive detections was statistically analyzed. A lower FPR indicates better robustness against false alarms in practical deployment scenarios. The FPR is calculated as follows:

$$FPR = \frac{FP}{FP + TN} \tag{16}$$

To ensure the validity and practical feasibility of the on-site experiment, it was divided into two phases: 1:1 simulation verification and on-vehicle installation. First, camera placement and angle optimization were performed on a 3D simulation model that meticulously replicated the structural parameters and operational environment of the LHD. This process involved iterative adjustments to maximize monitoring coverage in the front, rear, and side

directions while minimizing blind spots. The simulation phase was used to preliminarily verify the field-of-view overlap and the suitability of the camera-to-camera perspective configuration for a multi-view fusion strategy.

As shown in Figure 23, a total of five OAK-D depth vision cameras were installed on the LHD in a surround-view configuration. Two cameras were placed at the front of the LHD (installed at a height of approximately 2.4 m, symmetrically distributed to achieve a forward overlapping field of view), one was installed on each side (at a height of approximately 2.6 m to cover the lateral safety zones during LHD operation), and one was installed at the rear (at a height of approximately 1.5 m, specifically to cover the rear blind spot). Each camera captured video at a resolution of 1280 × 720 and a frame rate of 30 fps. The video streams were transmitted in real time via a wired channel to a T930G edge computing device (TWOWIN Technology, Shenzhen, China) installed on the LHD body. This device, equipped with an NVIDIA Jetson AGX Orin embedded GPU, was responsible for the parallel reception, preprocessing, and online inference of the multiple video streams. The edge device simultaneously deployed and ran both the baseline RT-DETR model trained in the lab and the proposed FP-DETR detection algorithm for real-time comparative evaluation and on-site verification. A unified timestamp and frame sequence management ensured the temporal consistency and fusion availability of the multi-camera detection results, thereby supporting real-time alerts, log recording, and subsequent multi-view fusion processing.

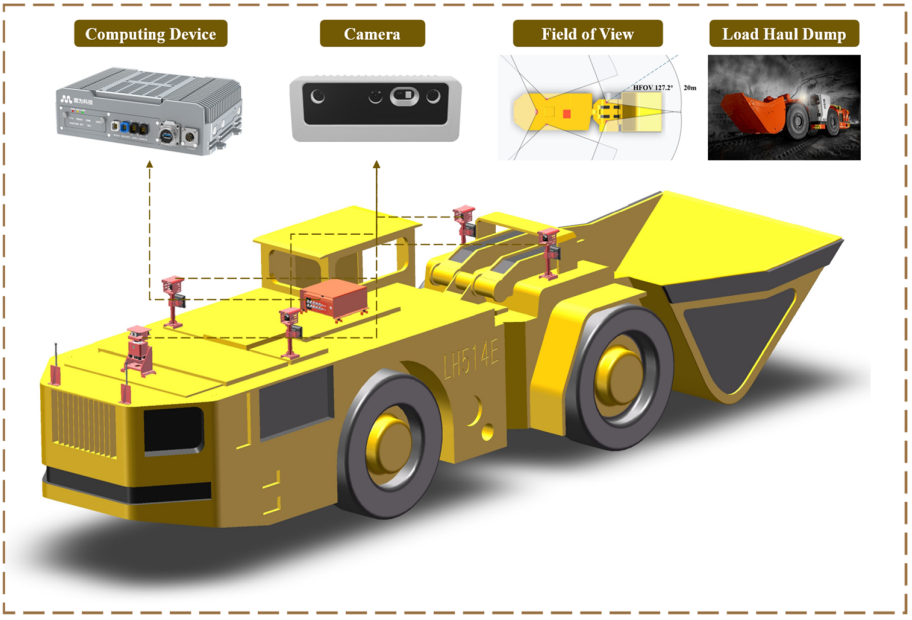


Figure 23. Schematic diagram of cameras mounted on an LHD in a simulated environment.

As shown in Figure 24, after determining the installation plan through simulation optimization, the system was physically installed on the actual LHD at the height and angle designed in the simulation. This approach prevented equipment obstruction while also meeting monitoring requirements in the actual working environment. A custom mechanical bracket was used to securely mount the camera to a predetermined position on the LHD body. The bracket was designed for easy angle adjustment and future maintenance. Industrial-grade, abrasion-resistant shielded cables were neatly routed along the LHD body, and all interfaces were protected to withstand the humid, dusty, and vibrating conditions underground.

(**a**) (**b**) (**c**)

Figure 24. Actual Camera Installation Locations On site. (**a**) Front; (**b**) Side; (**c**) Behind.

To improve equipment durability and measurement stability, the camera housing and mounting positions were equipped with dustproof, waterproof, and shock-absorption measures. Additionally, stress relief and protective treatments were applied at critical connection points to reduce vibration-related failure rates. The power and data links use industrial-standard power supplies and connectors to ensure long-term operational reliability. Live operational results show that this installation and deployment solution not only satisfies engineering constraints but also supports the practical validation of the proposed FP-DETR algorithm in real-world underground environments. All collected data were processed in real time on edge devices to compare the detection performance of the baseline RT-DETR-r18 and the proposed FP-DETR. The field test results show that FP-DETR achieves improved detection accuracy and a lower FPR under practical underground conditions. These results further demonstrate the performance and feasibility of the proposed engineering pipeline, including model optimization, edge-device deployment, and LHD-based field installation.

According to the results of the previous sections, the proposed FP-DETR demonstrates improved detection performance while maintaining reduced model complexity, and both mAP@0.5 and mAP@0.5:0.95 outperform the improved baseline. To analyze false alarm behavior, we additionally evaluate the frame-level FPR under scenarios without human presence, focusing specifically on no-person false alarm conditions. Here, FPR is defined as the proportion of frames incorrectly identified as containing personnel among all evaluated frames. A total of 17,648 frames were collected in the field test, and detection results were manually reviewed and annotated to compute the false alarms.

As shown in Table 10, RT-DETR-r18 produced 725 false alarms, corresponding to an FPR of 4.1%, while FP-DETR reduced the number of false alarms to 600, achieving an FPR of 3.4%, representing a relative reduction of approximately 17.1%. The FPS values reported in Table 10 are measured on a single-camera video stream deployed on an edge device, where FP-DETR achieves 65 FPS (approximately 15.4 ms per frame inference time). During deployment, the five camera streams were processed asynchronously on the edge device. The reported FPS corresponds to the average inference throughput per camera stream.

Considering the full pipeline including frame acquisition, preprocessing, inference, and alarm triggering, the estimated end-to-end latency is approximately 30–50 ms.

Table 10. Comparison of False Alarm Rate and Single-camera Inference Speed.

Model	Total Frames	False Alarms	FPR	FPS
YOLOv8m	17,648	782	4.4%	78
YOLO11m	17,648	882	5.0%	68
RT-DETR-r18	17,648	725	4.1%	42
FP-DETR	17,648	600	3.4%	65

It should be noted that the current field test is limited in scope, as it only evaluates no-person scenarios and focuses on false positive behavior. Due to the lack of personnel-present field testing, metrics such as true positives, false negatives, precision, recall, F1 score, detection delay, and missed-detection cases are not included in this study. Therefore, comprehensive safety performance under fully operational underground conditions cannot be fully established based on the current experiments.

Overall, the field experiment results suggest that FP-DETR can reduce false alarms while maintaining real-time performance under the tested conditions. However, its effectiveness for safety-critical deployment in underground mining environments should be interpreted cautiously, and further evaluation in scenarios involving both presence and absence of personnel is required to fully validate its practical reliability.

We recorded the detection results of the LHD (Load–Haul–Dump machine) during the field experiments and compared the detection result videos with annotations from manual on-site video samples. As shown in Figure 25, we selected several classic false alarm scenarios. The alarm mounted on the wall exhibits visual features like those of a safety helmet, causing both the baseline model and the improved model to mistakenly identify it as a person. However, you can see that even when the improved model made a false alarm, its confidence score was lower than that of the baseline model. In dynamic and complex scenarios, the baseline model had false detections when dealing with motion blur and complex lighting, whereas our improved model did not. These results demonstrate the improved stability of the proposed model in dynamic and complex underground environments. Regarding highly reflective objects, the baseline model often makes incorrect predictions when faced with glowing or reflective scenarios. This suggests that the model may struggle to distinguish subtle features during training, leading it to learn incorrect information. Our improved model, however, did not have any false alarms in these situations.

As shown in Figure 26, we conducted a heatmap comparison analysis for typical false detection scenarios in underground monitoring tasks. The figure shows that the baseline model's attention distribution is clearly skewed when faced with complex backgrounds. It tends to incorrectly focus on distracting objects that look like people, such as pipes on underground equipment, alarms mounted on mine walls, or shadows created by lighting. This unstable feature-focusing mechanism leads to a high false detection rate, which severely impacts the model's reliability in real-world applications. In contrast, our improved model demonstrates a more rational and concentrated attention distribution under the same conditions. Specifically, it has a stronger ability to distinguish between similar textures and distracting elements in the background. The heatmap results indicate that the improved model not only enhances its sensitivity to target features but also suppresses interference from irrelevant information, moderately reducing false alarms. Furthermore, the overall comparison shows that the improved model has higher stability and generalization in its feature extraction and discrimination capabilities. This advantage allows it to maintain

high stability even in extreme environments with low light and complex backgrounds. In summary, the heatmap analysis verifies that the improvements in the model's feature learning mechanism are effective, providing more robust performance than the baseline.

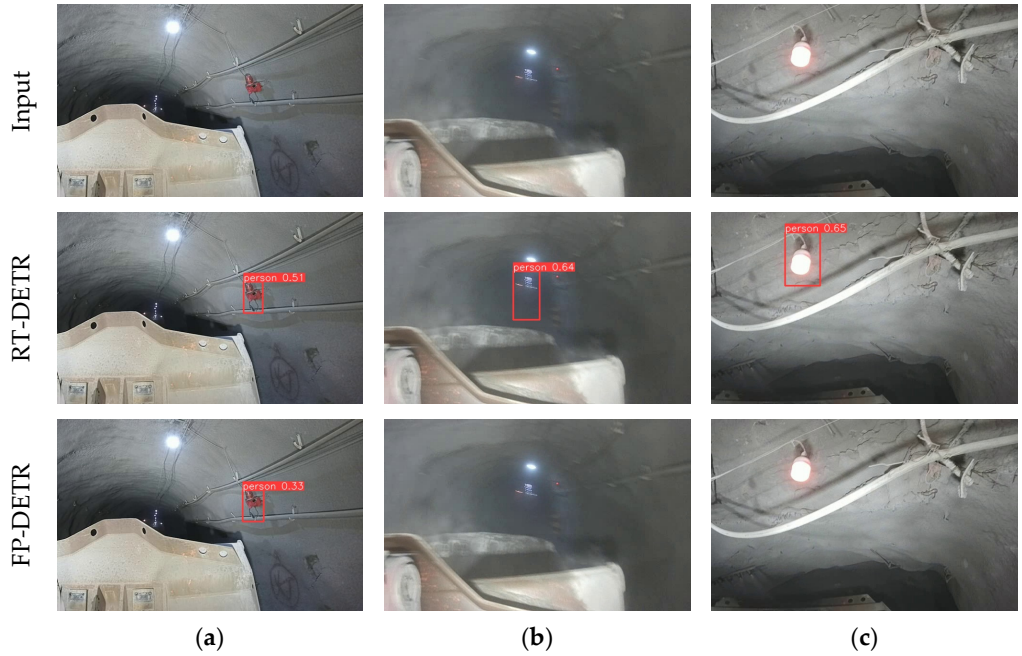


Figure 25. Visualization Examples of False Detection Results from On-Site Comparative Experiments. (**a**) Scene like hard hats; (**b**) Dynamic and complex scene; (**c**) High-brightness scene.

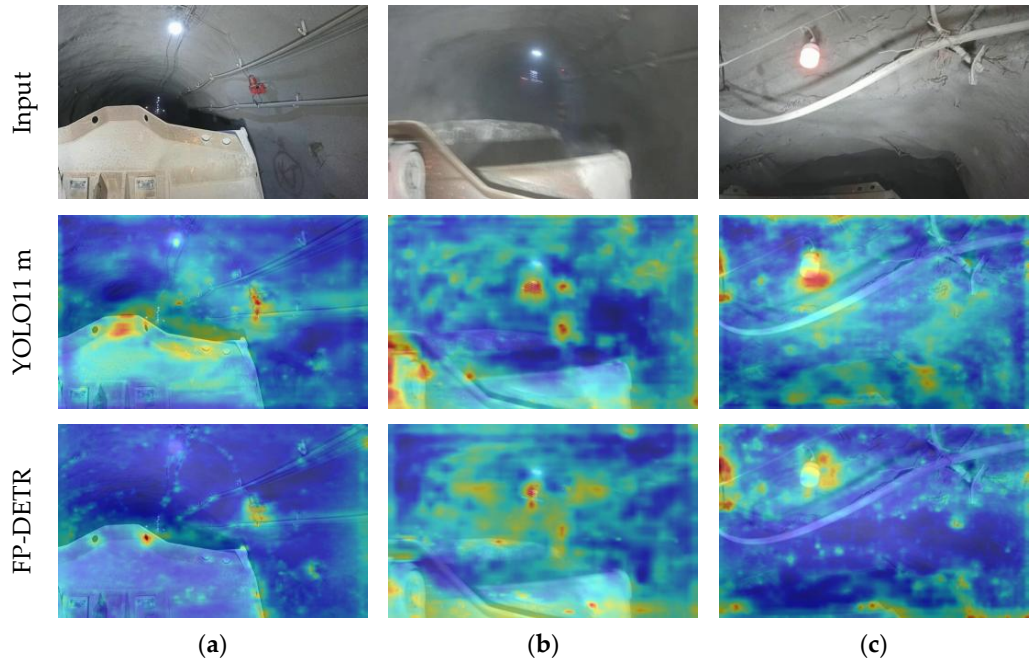


Figure 26. Detailed Heatmaps of False Detections in Different On-Site Experimental Scenarios. (**a**) Scene like hard hats. (**b**) Dynamic and complex scene; (**c**) High-brightness scene. The red area represents the area where attention is focused.

5.4. Discussion

This dataset presents several significant challenges for underground personnel detection. Personnel far from the mining equipment occupy only a small number of pixels, making accurate detection difficult. Occlusion caused by camera installation location and

mining equipment can further increase the false negative rate. Furthermore, miners' uniforms often have light textures and low contrast with the tunnel background, especially in dusty environments, causing personnel to blend into their surroundings. Uneven lighting, poor visibility, and low color contrast in underground tunnels also negatively impact detection robustness. Variations in camera resolution, installation location, and shooting angle also affect the sharpness and spatial distribution of personnel in images, leading to performance fluctuations across different scenarios.

Detection failures may occur at distance from the camera or in extremely low light and high dust environments. In some cases, due to similar visual features, the model may confuse background structures or equipment edges with personnel. Additionally, the current dataset was collected from limited underground environments and equipment configurations, which may limit the model's generalization ability under other mining or operational conditions. Future work will focus on improving small object detection capabilities, enhancing the model's robustness in low visibility conditions, and expanding the diversity of the dataset to improve the model's generalization ability.

6. Conclusions

This paper addresses the challenges of insufficient detection accuracy, high false alarm rates, and excessive computational complexity in underground personnel detection around Load–Haul–Dump (LHD) vehicles by developing an improved RT-DETR-based detection framework termed FP-DETR. Rather than introducing entirely new detection architectures, this work focuses on the moderate integration and adaptation of lightweight feature extraction, attention mechanisms, and loss optimization strategies for complex underground mining environments.

Specifically, the CSP-FFCM module is incorporated into the backbone network to enhance spatial–frequency feature representation under low-illumination and dust-interference conditions while reducing model complexity. In addition, the POTE module adapted from the polarity-aware linear attention mechanism in PolaFormer is integrated to improve global feature interaction and directional sensitivity with linear computational complexity. Furthermore, the Matching-Aware Loss (MAL) is adopted to improve the alignment between classification confidence and localization quality, thereby reducing false detections under severe foreground–background imbalance conditions.

Experimental results on the self-constructed underground dataset demonstrate that the proposed FP-DETR achieves consistent performance improvements over the baseline RT-DETR-r18 model. Specifically, FP-DETR improves mAP@0.5 from 87.0% to 88.5% and mAP@0.5:0.95 from 55.1% to 58.6%, while reducing the number of parameters from 20.0 M to 15.8 M and decreasing computational complexity from 57.3 GFLOPs to 51.2 GFLOPs. Multi-seed experiments further indicate the stability and reproducibility of the proposed framework under different training conditions. In field deployment experiments, FP-DETR also reduces the frame-level FPR from 4.1% to 3.4% compared with the baseline model, while maintaining real-time inference performance on edge devices.

Overall, the proposed FP-DETR achieves a reasonable balance between detection accuracy, computational efficiency, and deployment cost for underground personnel detection tasks. The experimental and field-test results suggest that the proposed framework has potential applicability for real-time underground perception systems. However, the current field validation remains limited to specific mining environments and deployment conditions. Future work will further investigate large-scale multi-mine validation, long-term deployment stability, and stability under more diverse underground operating scenarios.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, P.P. and K.C.; methodology, K.C.; software, P.P.; validation, C.Z.; formal analysis, X.L.; investigation, H.Z.; resources, P.P.; data curation, K.C.; writing—original draft preparation, K.C.; writing—review and editing, P.P.; visualization, S.G.; supervision, P.P.; project administration, P.P.; funding acquisition, P.P. and C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under grant 52374168, the National Key Research and Development Program of China under grant 2022YFC2904105 and 2023YFC2907403, the Science and Technology Innovation Program of Hunan Province under grant 2023RC3069, and the Fundamental Research Funds for the Central Universities of Central South University under grant 2025ZZTS0544.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the data collection involved non-invasive observation in workplace environments and no sensitive personal information was collected.

Informed Consent Statement: Informed consent was obtained from all photographed individuals involved in the study prior to data acquisition. Permission for on-site image and data collection was granted by the mining company and site management. All collected data were anonymized prior to analysis, and no personally identifiable information was disclosed in this study.

Data Availability Statement: The original contributions presented in this study are included in the article. For further inquiries, please contact the corresponding author. The source code and datasets will be deposited in the project repository and made publicly available upon publication. The repositories can be accessed at https://github.com/Keshawn082/FP-DETR/tree/main (accessed on 12 March 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LHDs	Load–Haul–Dump Vehicles
FP-DETR	DETR Detection with Fused Fourier Convolution and Polarity-aware Transformer Encoder
RT-DETR	Real-Time Detection Transformer
YOLO	You Only Look Once
CSP	Cross Stage Partial
FFCM	Fused Fourier-Conv Mixer
POTE	Polarity-aware Transformer Encoder
MAL	Matching-Aware Loss
P	Precision
R	Recall
AP	Average Precision
mAP	Mean Average Precision
FPS	Frames Per Second
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
FPR	False Positive Rate

References

1. Kia, K.; Fitch, S.M.; Newsom, S.A.; Kim, J.H. Effect of Whole-Body Vibration Exposures on Physiological Stresses: Mining Heavy Equipment Applications. *Appl. Ergon.* **2020**, *85*, 103065. [CrossRef]
2. Szrek, J.; Zimroz, R.; Wodecki, J.; Michalak, A.; Góralczyk, M.; Worsa-Kozak, M. Application of the Infrared Thermography and Unmanned Ground Vehicle for Rescue Action Support in Underground Mine—The AMICOS Project. *Remote Sens.* **2020**, *13*, 69. [CrossRef]
3. Shi, W.; Li, J.; Chen, Z.; Zhang, C.; Yang, H. Smart Mine: The Role of Artificial Intelligence in Modern Coal Extraction. In Proceedings of the 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Bangkok, Thailand, 9–12 December 2024; pp. 874–879.
4. Imam, M.; Baïna, K.; Tabii, Y.; Mostafa Ressami, E.; Adlaoui, Y.; Benzakour, I.; Bourzeix, F.; Hassan Abdelwahed, E. Ensuring Miners' Safety in Underground Mines Through Edge Computing: Real-Time PPE Compliance Analysis Based on Pose Estimation. *IEEE Access* **2024**, *12*, 145721–145739. [CrossRef]
5. Zhan, K.; Zhang, Y.; Ma, Z.; Liu, X.; Lv, X.; Liu, G.; Wang, H.; Liang, X. Key Technologies for Intelligent Mining of Underground Metal Mines. *MetaResource* **2024**, *1*, 13–22. [CrossRef]
6. Shahmoradi, J.; Talebi, E.; Roghanchi, P.; Hassanalian, M. A Comprehensive Review of Applications of Drone Technology in the Mining Industry. *Drones* **2020**, *4*, 34. [CrossRef]
7. Kim, H.; Choi, Y. Autonomous Driving Robot That Drives and Returns along a Planned Route in Underground Mines by Recognizing Road Signs. *Appl. Sci.* **2021**, *11*, 10235. [CrossRef]
8. Imam, M.; Baïna, K.; Tabii, Y.; Benzakour, I.; Adlaoui, Y.; Ressami, E.M.; Abdelwahed, E.H. Anti-Collision System for Accident Prevention in Underground Mines Using Computer Vision. In *Proceedings of the 2022 The 6th International Conference on Advances in Artificial Intelligence*; ACM: Birmingham, UK, 2022; pp. 94–101.
9. Patrucco, M.; Pira, E.; Pentimalli, S.; Nebbia, R.; Sorlini, A. Anti-Collision Systems in Tunneling to Improve Effectiveness and Safety in a System-Quality Approach: A Review of the State of the Art. *Infrastructures* **2021**, *6*, 42. [CrossRef]
10. Li, X.; Gu, Q.; Zhang, D.; Chen, L.; Xue, B. Rolling Clustering Algorithm for Open-Pit Mine Road Network Extraction Based on Vehicles Trajectory. *MetaResource* **2025**, *2*, 115–131. [CrossRef]
11. Zhang, J.; Yan, Q.; Zhu, X.; Yu, K. Smart Industrial IoT Empowered Crowd Sensing for Safety Monitoring in Coal Mine. *Digit. Commun. Netw.* **2023**, *9*, 296–305. [CrossRef]
12. Zhao, A.; Zheng, Q.; Zhao, Y.; Li, L. MSC-DETR: A Lightweight Real-Time Foreign Object Detection Algorithm for Underground Belt Conveyors in Coal Mines. *J. King Saud Univ. Comput. Inf. Sci.* **2026**, *38*, 116. [CrossRef]
13. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef]
14. Imam, M.; Baïna, K.; Tabii, Y.; Ressami, E.M.; Adlaoui, Y.; Benzakour, I.; Abdelwahed, E.H. The Future of Mine Safety: A Comprehensive Review of Anti-Collision Systems Based on Computer Vision in Underground Mines. *Sensors* **2023**, *23*, 4294. [CrossRef]
15. Yang, W.; Zhang, X.; Ma, B.; Wang, Y.; Wu, Y.; Yan, J.; Liu, Y.; Zhang, C.; Wan, J.; Wang, Y.; et al. An Open Dataset for Intelligent Recognition and Classification of Abnormal Condition in Longwall Mining. *Sci. Data* **2023**, *10*, 416. [CrossRef]
16. Xu, P.; Zhou, Z.; Geng, Z. Safety Monitoring Method of Moving Target in Underground Coal Mine Based on Computer Vision Processing. *Sci. Rep.* **2022**, *12*, 17899. [CrossRef]
17. Wei, X.; Zhang, H.; Liu, S.; Lu, Y. Pedestrian Detection in Underground Mines via Parallel Feature Transfer Network. *Pattern Recognit.* **2020**, *103*, 107195. [CrossRef]
18. Wang, W.; Wang, S.; Zhao, Y.; Tong, J.; Yang, T.; Li, D. Real-Time Obstacle Detection Method in the Driving Process of Driverless Rail Locomotives Based on DeblurGANv2 and Improved YOLOv4. *Appl. Sci.* **2023**, *13*, 3861. [CrossRef]
19. Shao, X.; Liu, S.; Li, X.; Lyu, Z.; Li, H. Rep-YOLO: An Efficient Detection Method for Mine Personnel. *J. Real-Time Image Proc.* **2024**, *21*, 28. [CrossRef]
20. Wang, W.; Hu, K.; Jiang, H. Multi-Target Detection Method for Driving Area of Mine Driverless Rail Locomotives Based on Lightweight Network. *SIViP* **2025**, *19*, 276. [CrossRef]
21. Zijian, T.; Kang, Y.; Jiaqi, W.U.; Wei, C. LMIENet enhanced object detection method for low light environment inunderground mines. *Coal Sci. Technol.* **2024**, *52*, 222–235. [CrossRef]
22. Tian, F.; Wang, M.; Liu, X. Low-Light Mine Image Enhancement Algorithm Based on Improved Retinex. *Appl. Sci.* **2024**, *14*, 2213. [CrossRef]
23. Ni, Y.; Huo, J.; Hou, Y.; Wang, J.; Guo, P. Detection of Underground Dangerous Area Based on Improving YOLOV8. *Electronics* **2024**, *13*, 623. [CrossRef]
24. Jin, H.; Ren, S.; Li, S.; Liu, W. Research on Mine Personnel Target Detection Method Based on Improved YOLOv8. *Measurement* **2025**, *245*, 116624. [CrossRef]

25. Li, Y.; Yan, H.; Li, D.; Wang, H. Robust Miner Detection in Challenging Underground Environments: An Improved YOLOv11 Approach. *Appl. Sci.* **2024**, *14*, 11700. [CrossRef]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
27. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12346, pp. 213–229, ISBN 978-3-030-58451-1.
28. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs Beat Yolos on Real-Time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974.
29. Song, Y.; Qian, P.; Zhang, K.; Liu, S.; Zhai, R.; Song, R. An Improved Multi-Scale Fusion and Small Object Enhancement Method for Efficient Pedestrian Detection in Dense Scenes. *Multimed. Syst.* **2025**, *31*, 151. [CrossRef]
30. Tian, F.; Song, C.; Liu, X. Small Target Detection in Coal Mine Underground Based on Improved RTDETR Algorithm. *Sci. Rep.* **2025**, *15*, 12006. [CrossRef]
31. Zhang, S.; Benenson, R.; Schiele, B. Citypersons: A Diverse Dataset for Pedestrian Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 3213–3221.
32. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761. [CrossRef] [PubMed]
33. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2014; Volume 8693, pp. 740–755, ISBN 978-3-319-10601-4.
34. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking Single-Image Dehazing and Beyond. *IEEE Trans. Image Process.* **2019**, *28*, 492–505. [CrossRef] [PubMed]
35. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
36. Chu, T.; Chen, J.; Sun, J.; Lian, S.; Wang, Z.; Zuo, Z.; Zhao, L.; Xing, W.; Lu, D. Rethinking Fast Fourier Convolution in Image Inpainting. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Paris, France, 2023; pp. 23138–23148.
37. Gao, N.; Jiang, X.; Zhang, X.; Deng, Y. Efficient Frequency-Domain Image Deraining with Contrastive Regularization. In *Computer Vision—ECCV 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2025; Volume 15099, pp. 240–257, ISBN 978-3-031-72939-3.
38. Zu, Y.; Li, S.; Zhang, L. POLAR-DETR: Polarized Occlusion-Aware Local-Global Attention Real-Time Detection Transformer for Total Laboratory Automation. *Sci. Rep.* **2026**, *16*, 11949. [CrossRef] [PubMed]
39. Meng, W.; Luo, Y.; Li, X.; Jiang, D.; Zhang, Z. Polaformer: Polarity-Aware Linear Attention for Vision Transformers. *arXiv* **2025**, arXiv:2501.15061.
40. Peng, Y.; Li, H.; Wu, P.; Zhang, Y.; Sun, X.; Wu, F. D-FINE: Redefine Regression Task of DETRs as Fine-Grained Distribution Refinement. In Proceedings of the International Conference on Learning Representations, Singapore, 24–28 April 2025; Volume 2025, pp. 44015–44031.